MAURICON

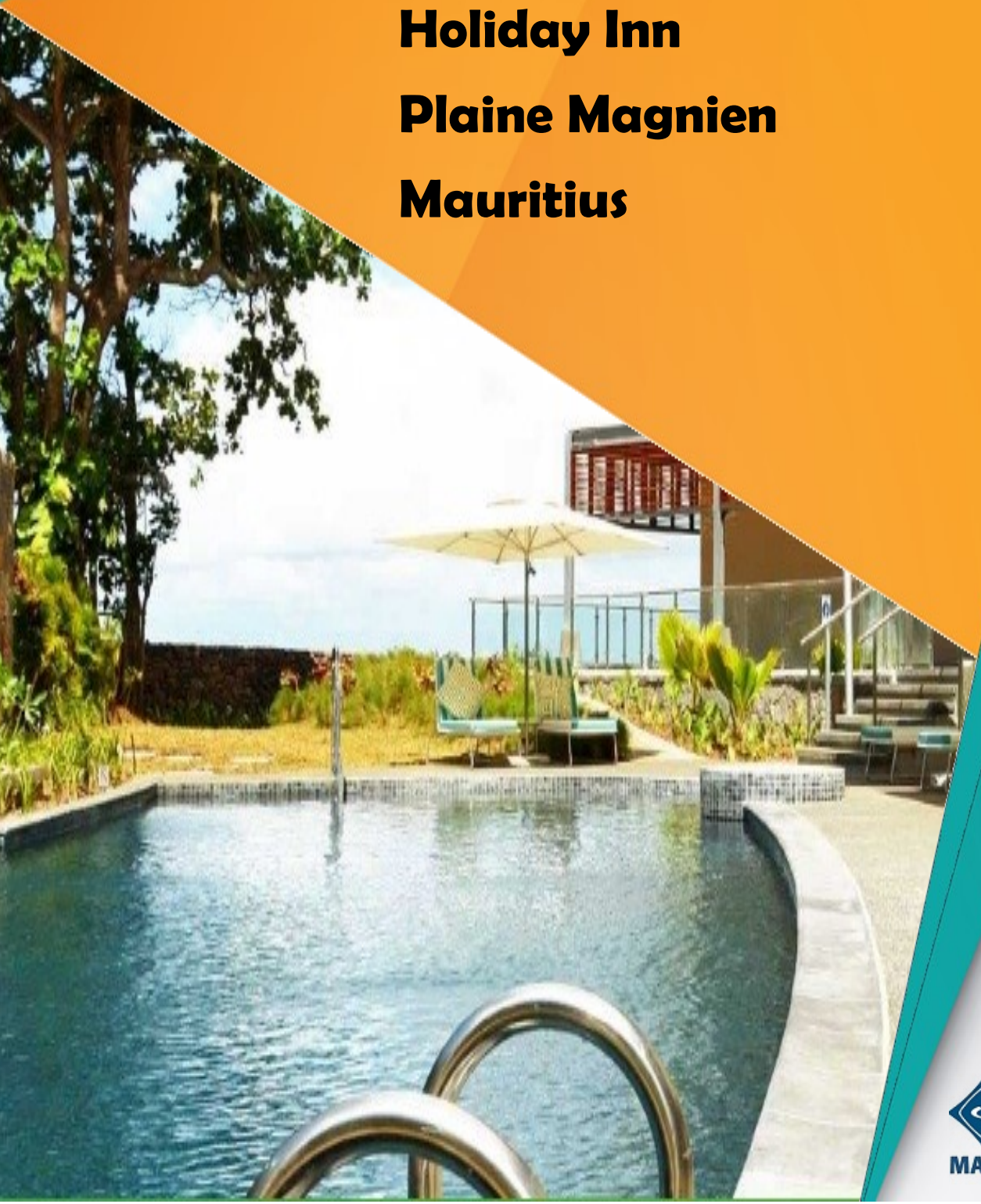
2018 International Conference on Intelligent & Innovative Computing Applications (ICONIC)

6th & 7th December 2018

Holiday Inn

Plaine Magnien

Mauritius



SITE©

IEEE
MAURITIUS SUBSECTION

Copyright

2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC)

Part Number CFP18P38-USB

ISBN 978-1-5386-6476-6

General Chair

Professor (Dr) Krishnaraj Madhavjee Sunjiv Soyjaudah, Senior Member of IEEE was the General Chair and the TPC Chair of AFRICON 2013. He received his BSc (Hons) degree in Physics from Queen Mary College, University of London in 1982 as a recipient of the UK commonwealth scholarship, MSc degree in Digital Electronics from King's college, University of London in 1991 again as a recipient of the UK Commonwealth Scholarship, PhD degree and Gold Medal for pure research from University of Mauritius in 1998. He then pursued legal studies and obtained a LLB (Hons) from University of London in 2007, Post Graduate Diploma in legal studies at the Bar from Manchester law school in 2015 as a recipient of the Vice Chancellor's Scholarship and called to the degree of the Utter Bar of the Honourable Society of Middle Temple Inn in November 2015.

As a Professor of Communication Engineering, he has served as Dean of the Faculty of Engineering at the University of Mauritius. Professor Soyjaudah has also held senior management positions in regulatory authorities namely as Executive Director of the Tertiary Education Commission and Executive Director of the Information and Communications Technologies Authority. As an academic researcher, Professor Soyjaudah has published over 240 refereed International Journal and Conference Research papers and has successfully supervised two post-doctoral fellows and 17 PhDs. He is currently supervising 12 PhD students.

His expertise is also solicited by foreign universities and he examined several PhDs from South African and Indian Universities. Professor Soyjaudah was the Chairman of the Mauritius Qualifications Authority from 2002 to 2005, Board Director of the Multicarrier Mauritius Limited (MCML) from 2001 to 2016 and chaired MCML's staff and finance and procurement committees. He served as Board Member, in the capacity of a technical expert, in the Energy Efficient Management Office from 2011 to 2015.

Message from the General Chair

The 2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC) will be held on the 6th and 7th of December 2018 at the Holiday Inn Mauritius Mon Tresor, in the paradise island of Mauritius. This conference is being organised by the Society of Information Technologists and Entrepreneurs with support from the IEEE Mauritius Subsection in order to provide an opportunity to academic researchers, industry professionals, government delegates and students to interact and share their experiences and knowledge on cutting-edge developments in the fields of Cloud Computing, Internet-of-Things, Networked Systems, Data Communications, Computer Security, Multimedia, Computer Vision, Intelligent Computing, Robotics, Machine Vision, Signal Processing, Software Engineering, ICT4D and ICT in Education. Our main objective is to promote scientific and educational activities towards the advancement of knowledge by improving the theory and practice of various disciplines and areas of ICT Engineering.

Papers are invited for Mauricon 2018 on topics lying within the scope of the conference. All contributions must be of high quality, original, and must not have been previously published elsewhere or intended for publication elsewhere. All papers will be reviewed by members of the International Programme Committee and depending on their degree of innovativeness and quality of presentation, may be selected for oral or poster presentation and publication in the conference proceedings. All accepted and registered papers will be submitted for potential inclusion in IEEE Xplore Digital Library (Conference Number #44423), provided they meet IEEE quality standards. All submissions will be reviewed by at least two independent reviewers by following a double-blind review process.

We look forward to welcoming you to Mauritius in December 2018 so that you can experience the Mauritian hospitality, culture and atmosphere.

Keynote Speakers

Dr Siamak Solat is currently working as a Research Engineer of Blockchain and Distributed Ledger Technology at Computer Science and Artificial Intelligence Laboratory of Engie, France. His responsibility is to find solutions for challenges and limitations of utilizing the Blockchain Technology in particular in energy trading. Earlier, he worked at the Laboratory of Computer Science of Paris 6 (LIP6) in France. This is a joint research laboratory of Université Pierre et Marie Curie (UPMC) and the French National Centre for Scientific Research (CNRS). He received his Master's degree in Computer Science and Networking from the Sorbonne University (France) from where he also received his PhD degree in Computer Science and Information Security. His PhD involved the assessment of the security of the blockchain technology and decentralised crypto-currencies. He proposed an innovative timestamp-free algorithm to prevent the Selfish Mining (block-withholding) attack in the Bitcoin network. He also proposed an alternative to the famous consensus mechanism, proof-of-work, to achieve a more democratic, fairer and energy efficient consensus for decentralized crypto-currencies named RDV: Register, Deposit, Vote.

His speech will be on the Blockchain Technology and Crypto-currencies.

Dr Anand Nayyar hold a Ph.D. in Computer Science in the area of wireless sensor networks and swarm intelligence. He is currently working as Professor of Computer Science at the Graduate School at the Duy Tan University in Da Nang, Vietnam. He has published extensively in various national and international conferences. He has also published several books in different areas of computer science. He is an editorial board member of many journals and has reviewed thousands of papers for conferences and journals. He is regularly invited to deliver keynote speeches by various universities, societies and committees. His current fields of interests include: wireless sensor networks, mobile adhoc networks, swarm intelligence, cloud computing, network security, wireless communications, Internet of Things (IoT), distributed computing and big data analytics. He is also a certified professional in CCNA, MCSE and Oracle certifications, amongst many others.

His speech will be on the Blockchain Revolution.

Organising Chairs

Dr. Upasana Singh is a lecturer in the Discipline of Information Systems and Technology at the University of KwaZulu Natal (UKZN) Westville Campus in Durban, South Africa. She lectures on a wide-range of IT-related subjects and she has a keen interest in Educational Technologies. Her PhD was on electronic assessment for which she developed and evaluated an online interactive tool to evaluate e-assessment systems, named e-SEAT. Using the e-SEAT framework academics can make a more informed decision on implementing the best-fit e-assessment system to satisfy their individual requirements. In 2014 and 2015, she received 3 research awards, one from the South African National Research Foundation (NRF) and the other from UKZN, College of Law and Management Studies (CLMS) Teaching and Learning Unit and the third from the UKZN Teaching and Learning Office (UTLO) for advancing her research and for her contributions to Teaching and Learning. She serves annually as panel reviewer for the NRFs Joint Collaboration applications since 2015. Joint academic collaboration is one of her fortes having secured 3 MoUs with international universities since 2016. This has provided her with the platform to share her knowledge as invited guest speaker to local and international conferences in South Africa, India and Mauritius since 2015.

Sameerchand Pudaruth is a Senior Lecturer at the Department of ICT in the Faculty of Information, Communication & Digital Technologies, at the University of Mauritius. He holds a First Class degree in Computer Science and Engineering and a Masters in Computer Science. In 2013, he graduated with an LLB from the University of London. His research expertise and interests are in Machine Learning and Artificial Intelligence. He has presented many research papers in international conferences and published many papers in internationally recognised journals. He is a member of the internationally renowned ACM and IEEE societies. He is an executive member of the IEEE Subsection of Mauritius. He is an MQA (Mauritius Qualifications Authority) approved trainer and has successfully delivered many training courses (Web Technologies, Mobile Applications Development, etc) on a national basis. Sameerchand is also the author of a programming book in Python: Python in One Week.

Technical Program Committee

Dr Vinaye Armoogum, University of Technology, Mauritius (Chair)
Professor Jan Haase, Helmut-Schmidt University of the Federal Armed Forces, Germany
Professor Wagacha Peter Waiganjo, University of Nairobi, Kenya
Professor Hein Venter, University of Pretoria, South Africa
Professor Marijke Coetzee, University of Johannesburg, South Africa
Professor Elize Ehlers, University of Johannesburg, South Africa
Professor Aurona Gerber, University of Pretoria, South Africa
Professor Philip Machanick, Rhodes University, South Africa
Professor Barry Irwin, Rhodes University, South Africa
Professor Yirsaw Ayalew, University of Botswana, Botswana
Professor Anand Nayyar, Duy Tan University, Vietnam
Dr. Eng. Hamid Ali Abed Al-Asadi, Basra University, Iraq
Dr Duronke Eyitayo, University of Botswana, Botswana
Dr Patricia Lutu, University of Pretoria, South Africa
Dr Amar Kumar Seeam, Middlesex University, Mauritius
Dr Nawdha Thakoor, University of Mauritius, Mauritius
Dr Keith Langenhoven, University of the Western Cape, South Africa
Dr Clement Nyirenda, University of Namibia, Namibia
Dr Xavier Bellekens, Abertay University, United Kingdom
Dr Peter Nabende, Makerere University, Uganda
Dr Nour Eldeen Mahmoud Khalifa, Cairo University, Egypt

Publication Committee

Sameerchand Pudaruth, Mauritius
Jeet Gohee, Mauritius

Registration and Local Organizing Committee

Somveer Kishnah, Mauritius
Rajiv Rughoobur, Mauritius

Table of Contents

Sustainability Simulation and Assessment of Bicycle Network Design and Maintenance Environment

Egils Ginters, Mareks Mezitis
Riga Technical University
Latvia
egils.ginters@rtu.lv

Dace Aizstrauta
Sociotechnical Systems Engineering Institute
Latvia
dace.aizstrauta@gmail.com

***Abstract*—Humans are an integral part of the nature and to avoid imbalance, the impact of technology on the surrounding environment should be limited. Therefore, green technologies are becoming an important part of economy, politics and science. One of the largest sources of pollution are motorized vehicles, that account for around 14 % of the global greenhouse gas emissions [1]. Use of green transportation within the integrated multimodal transport system is important to minimize carbon emissions [2]. One of these means of transportation is cycling. For cycling to become a fully integrated element within a multimodal transport system and not just a type of tourism, an appropriate infrastructure is necessary – cycling routes, lighting, bike rental, parking, repair services, as well as alignment with other types of transport. Cyclists need route planning tools, information about road surface quality, relief, and usage patterns during different weather conditions and over different days of the week. Infrastructure development projects demand significant financial resources and therefore municipalities need sustainable tools for services design and management. In this article the authors discuss acceptance and sustainability assessment methodology IASAM use to validate the VeloRouter - cycling network designing technology.**

Keywords—sustainability assessment, green transport, cycling network design and maintenance, simulation

I. Introduction

FP7-ICT-2011-7 FUPOL project No. 287119 (2011-2015) "Future Policy Modelling" [3] gave the opportunity to perform market research and understand the needs of potential users, interested in sustainable transport scheme development. Popularity of cycling in European cities is undoubtedly increasing. For example, in Helsinki more than 44 % of inhabitants are pedestrians or cyclists [4]. Besides according to data from European Cycling Federation [4], in several other EU capital cities, like Copenhagen, the proportion of cyclists can reach 30 % and more. The market analysis involves review of different products usable for route planning [5].

Nevertheless, it was concluded that existing products aimed at cycling routes design mainly offer the capabilities of planning and publishing the routes, but do not provide functionality necessary for municipalities to build justified bicycle path network. Skopje bicycle inter-modality simulator was designed under the framework of FUPOL project [3]. It was designed to find a usable solution for bicycle stations location in the City of Skopje. The system assisted the municipality to improve the scheduling and resource planning. However, simulation algorithms and results were focused on the needs of a large municipality rather than cyclists therefore wide usability of the technology is limited.

II. VeloRouter - Bicycle Network design and maintenance Environment

Urban transport system is multi modal (see Fig. 1) and includes several means of transportation that, in turn, defines certain interoperability demands. Therefore, the development of cycling paths cannot be voluntary. It is of high importance to acquire the opinion of cyclists about the most popular routes and the impact of other means of transportation. It is also important to build and maintain cycling routes within the boundaries of demands of citizens and municipal budgetary capabilities.

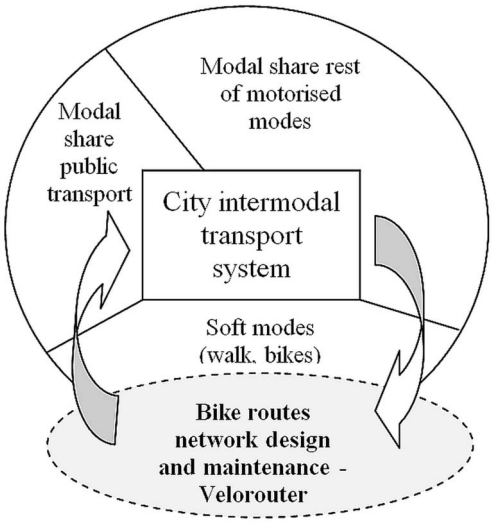


Fig. 1. VeloRouter in intermodal urban transport scheme.

Multi-agent based bicycle path network designing and exploitation simulator VeloRouter [5, 6] has dual applicability as it provides results for both audiences: municipalities and cyclists. By using VeloRouter municipalities could determine the eventually most popular routes, as well as interact and receive feedback from cyclists about the quality of paths. In

turn, cyclists could find out the potential occupancy of the route on the preferred day.

To generate data on each route's usage, the system uses agent-based simulation model in *Repast* environment and *Ingres* database storing data about routes use and intentions. Data are stored, and their obsolescence, credibility and eventual change of these routes are monitored. The sources are interviews with the cyclists and social media data. Route usage is simulated in macro, mesoscopic and micro level (see Fig. 2). Modeling results empowers municipalities to determine the priorities regarding cycling route development.

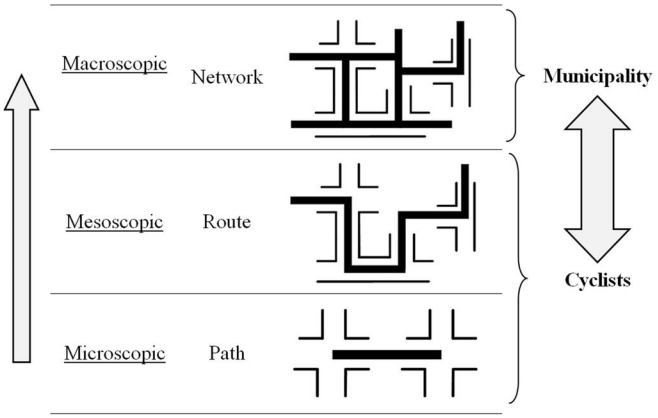


Fig. 2. Traffic simulation levels in VeloRouter.

In fact, VeloRouter has twofold functions – it is intelligent cycling network design and maintenance tool, as well as an e-citizen environment that enables communication with cyclists to hear out their preferences.

VeloRouter is a typical new generation simulator whose architecture provides both its design and its operation at the same time. User requirements are received through both direct requests and semantic analysis of social networks. Similarly, changes made to modeling algorithms are also validated. The basic element of the simulator is the intention and traveling database, which is automatically updated to prevent data aging. The longer the simulator is operated, the more current data is available, the more precise is the realization of route loading forecasting and network planning.

VeloRouter is an open system that enables the use of the data collected for more detailed analysis using semantic analytic and visualization, for example in *SemaVis* [7] environment. The purpose of diverse visualization is to bring the representation of the user's perceptions and habits of use closer. Semantic visualization enables not only to present the current situation, but also the possible causes that have affected the results achieved. It was tested during Skopje bicycle map modeling.

In VeloRouter also distributed simulation architectures as ECE [8] and HLA [9] can be used to ensure interoperability in integrated multilevel model.

However, implementation and exploitation of any technology demand significant financial resources, therefore the technology must be sustainable to attract investments. New technologies are emerging as a product of interaction between engineering and social sciences. If the social factor is not respected, the technology will be destined to fail. Will the society appreciate and accept the new technology? How sustainable the technology should be for the investment to become profitable? Is the implementation of VeloRouter expedient? Analogue questions may rise also regarding other existent and new technologies, their implementation and effectiveness.

III. SYSTEM DYNAMICS AND SKYPE BASED INTEGRATED ACCEPTANCE AND SUSTAINABILITY ASSESSMENT SIMULATION METHODOLOGY

Technology sustainability and development research was launched by authors in 2010 under the umbrella of FP7-ICT-2009-5 CHOREOS [10] project No. 257178. It turned out that there were practically no tools or assessment methodologies for technology acceptance and sustainability integrated evaluation.

Technology life-cycle approach concentrates on universal stages that can be applied to innovation research. This approach focuses rather on market forces and management decisions. In the literature, it is common to see the terms industry life cycle, product life cycle and technology life cycle used interchangeably, ambiguously and often inappropriately. Moreover, the discourse is dominated by the product life cycle while the technology life cycle has largely been neglected [11]. The technology life cycle is broader than a specific product or a process innovation, which applies to products sold in different markets [12]. Taylor&Taylor [12] pointed out that this is only the tip of the iceberg since there are also disconnects and inconsistencies pertaining to the various perspectives on the technology life cycle. This approach does not answer the questions regarding the technology acceptance and sustainability as it concentrates rather on commercial/managerial problems and views technology as separate item and does not analyse the differences of the technologies themselves.

Approaches such as diffusion of innovations, Technology Acceptance Model (TAM) or Unified Theory of Acceptance and Use of Technology (UTAUT) [13, 14, 15] focus on technology adoption and technology acceptance from individual's perspective during the decision-making process. First, these theories concentrate on the reasons behind the initial decision to use a technology, not the decision to continue to use it. Second, they are oriented towards existing, fully developed technologies, and involved the need for potential users to verify the technology. These theories do not involve integrated assessment of both, social and technological factors, that is an integral part of sociotechnical systems acceptance and sustainability evaluation. These

theories should be perceived as theoretical frameworks that illustrate and help to understand the decision itself but cannot be applied in fast decision making and technology sustainability evaluation.

The first concept of Integrated Acceptance and Sustainability Assessment Model (IASAM) was elaborated for evaluation the CHOReOS project results [16, 17] (see Fig. 3).

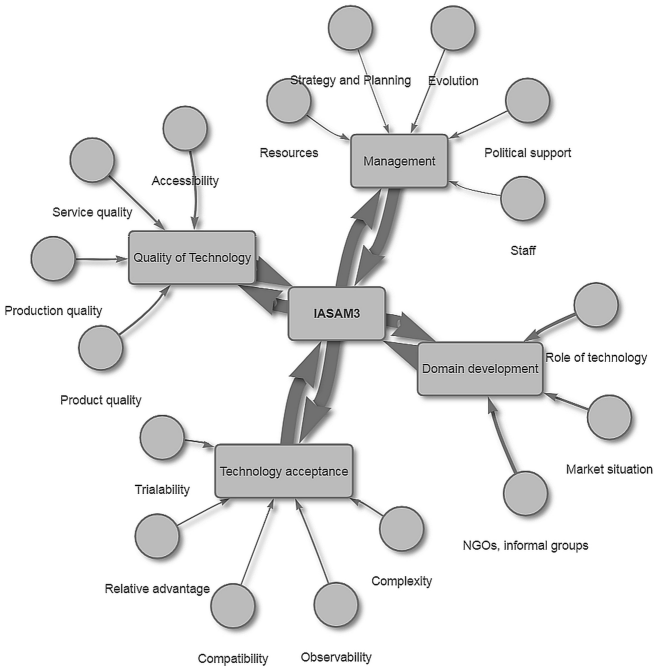


Fig. 3. System dynamics simulation based IASAM conceptual model.

Developed system dynamics model consists of four basic data flows: Management, Quality, Domain development and Acceptance. The interoperability of the flows are simulated in custom modified *InsigthMaker* [18] environment thus changing the IASAM index and measuring sustainability as the common resource.

The changes of sustainability index of *i*-technology $IASAM(Sust_T^i)$ respecting feedbacks are described at macro level by the equation (1):

$$IASAM(Sust_T^i(t)) = IASAM(Sus_T^i(t-dt)) + \\ +(Manage_T^i + Quality_T^i + Domain_T^i + Accept_T^i) * dt \quad (1)$$

The value of each flow is calculated from values of 59 parameters, furthermore, each parameter and each flow have its own weight within the total sustainability index calculation.

IASAM is a self-assessment methodology. Management flow describes the quality of management of the project/technology. Of course, no special investments in project management are needed in case of short term mobile application, but for long term or financially intensive projects agile approach might involve too high risks. Quality flow evaluates the clearness and transparency of the technology, thus assessing, whether it can be introduced and managed. The Domain development flow is very important as it includes assessment of social dimension, specifically the impact of societal, nongovernmental and market activities. As well Acceptance flow looks at technology acceptance prospects.

The values of the important parameters are assigned in conformity with Likert seven-point scale.

IASAM is a useful tool for the what-if scenario analysis and easy to customize the needs of experts. Each expert can change the flow weight affecting the sustainability in integrated solution. Initially, the weight of each flow is equal to the 25%. However, if the examiner considers that the proportion should be different, then necessary changes using the sliders can be conveniently implemented. During assessment session described in the current research it is assumed that all flow impact is equally important.

The framework of IASAM envisions the usage of Skype life-cycle curve for prediction of technology sustainability and uses *skypes* to express the IASAM sustainability index. This style of rating provides good comprehensibility, but the use of web tools gives easy access for all shareholders. IASAM index breakdown consists of four groups [0.00-0.25), [0.25-0.50), [0.50-0.75) and [0.75-1.00]. Each group reflects sustainability level of the assessed technology.

1) [0 – 0,25): IASAM criteria have received low evaluations. Many of the issues regarding the management, quality of technology, acceptance and domain development are inadequately performed or deliberated. Technology poorly satisfies the defined IASAM criteria;
2) [0,25 - 0,5): IASAM criteria have received low to medium evaluations. Although some issues are worked out, there are still many issues that remain unresolved regarding the management, quality of technology, acceptance and/or domain development. Technology somewhat satisfies the defined IASAM criteria, but the acceptance and sustainability of the technology is doubtful;
3) [0,5 – 0,75): IASAM criteria have received medium to high evaluations. Important issues regarding the management, quality of technology, acceptance and domain development are adequately performed or deliberated. Technology on average satisfies the defined IASAM criteria well. The acceptance and sustainability of the technology is promising, and if the work on technology development is carried on and the issues that gained lower evaluation are kept in close watch, the technology might become widely accepted (among target group) and sustainable in terms of existence of technology, financial gain and other targets set by its developers;
4) [0,75 - 1]: IASAM criteria have received high evaluations. Important issues regarding the management, quality of technology, acceptance and

domain development are fully performed or deliberated. Technology highly satisfies the defined IASAM criteria. The technology proves to be accepted among target group members and sustainable in terms of existence of technology, financial gain and other targets set by its developers.

There could be multiple iterations of each project appraisal, particularly useful for self-evaluation of the project/technology, because changing the values of influencing parameters promotes finding the best implementation and introduction variant.

IV. VeloRouter Sustainability Assessment case

VeloRouter simulator was validated as a new technology that can be used for sustainable bicycle routes network intelligent development and managing.

The assessed values of VeloRouter technology influencing flows achieve the following: Management flow - 86% of maximum, Quality flow - 94%, Domain development flow - 93%, but Technology acceptance - 97%. Pay attention that the results are not statistical but self-assessment simulated by IASAM.

The analysis of the Management flow showed that the team is technically good equipped, has the necessary knowledge and funding. It has evaluated competitors and the demands of the target group; the risks and available resources are being managed. The developers are experienced, but there may be doubts about the motivation of staff. From the political point of view, this technology is neutral, it is not expected to meet resistance due to ethical problems. VeloRouter technology embeds society's feedback accounting and risks management mechanisms.

The analysis of the Quality flow revealed that technology has clearly defined result/output, and it can be adequately disseminated. The product does not demand specific equipment, and the user should have only basic computer knowledge.

Domain development review looks at the impact of informal groups and nongovernmental organizations, as well as the market situation. There are practically no analogous tools to VeloRouter, therefore it can be positioned as a niche product.

Acceptance measures user's willingness to accept the technology. So, the model guess that potential audience is ready for introduction.

VeloRouter could give instantaneous economic benefits and its implementation costs are low. The product meets customer expectations and does not contradict with socio-cultural environment and traditions. VeloRouter is easy to use, it provides demo version and user manuals. In addition, the output of the technology is visually comprehensible. So, it would be appreciated by society. Such suggestions are confirmed by IASAM.

VeloRouter sustainability index $IASAM(Sust_T^i)$ self-assessment result is 0.92 *skypes* [0.75-1.00]. This proves the stability and sustainability of the cycling network designing and management technology (see Fig. 4).

Important issues regarding the management, quality of technology, acceptance and domain development are fully performed or deliberated. VeloRouter technology highly satisfies the defined IASAM criteria. The technology proves to be accepted among target group members and is sustainable in terms of potential existence of technology, financial gain and other targets set by its developers.

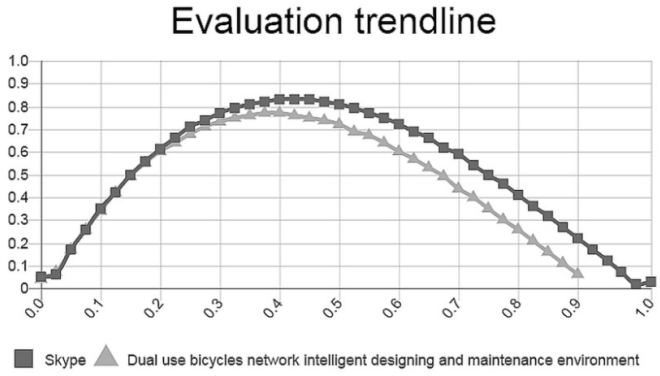


Fig. 4. VeloRouter sustainability index curve by IASAM.

The assessment was carried out also by many experts that constitute the statistically significant result with confidence level 0.95. The result of Kolmogorov-Smirnov confidence test value for VeloRouter assessment was 0.85 and did not exceed the critical value. It shows the homogeneity of the evaluation results and their suitability for reasoned conclusions.

Assessment results of VeloRouter showed that technology is sustainable and can be implemented successfully. In addition, this tool could become an open and intelligent component of design and management of municipality's multimodal transportation system.

V. Conclusions

Real sociotehnical system involves technical infrastructure and the social component. It defines the broad set of important and influential parameters, as well as permanent changes and evolution of the system, and stochastic influence on both the implementation and maintenance.

The simulation can be used in transport scheme for the elements interoperability and changes modeling that allows what-if scenario analysis at different levels of the hierarchy.

Sustainabilities forecast quality is determined by the model validation results and operational data reliability. However, the confidence of the data depends on the data aging, the accumulated volume of the data, representation of the audience and plausibility of the views expressed. Therefore, social networking is integral attribute of the

VeloRouter operation. The entrance for VeloRouter customers are social networks as LinkedIn, Facebook etc.

An important attribute of modern technology in accordance with Finance 4.0 paradigm is indirect and additional benefit to the provided service. VeloRouter provides load simulation in the selected route. This gives the ability to plan the time and to promote the cycling aimed to reducing traffic and pollution in the city.

IASAM methodology used for VeloRouter sustainability assessment is among rare quantitative approaches that enable assessment of new and existing technology sustainability and can be used in investment decision process.

The third version of the IASAM [19] is significantly simplified, and the procedure is made less laborious to promote self-assessment capabilities. The main advantages of the IASAM methodology are simplicity and accessibility while providing robust quantitative results.

VeloRouter assessment by IASAM is sufficiently credible and confirms the sustainability of the dual use technology, as well as justify its suitability for urban green transport planning and operation. The above approach confirms that the future is diverse products that can be used at the same time at different levels of organizational structure.

Further work related with VeloRouter development will involve expanding of planning region, that will inevitably increase the demand of computing power and asking for simplification of simulation. The agent-based simulation model now has to remember the possible location of any bike on a common map. Simulation algorithms will be simplified by object grouping. The purpose of the enhancements is to provide route planning and evaluation capabilities to the smart phone user. The authors also intend to launch the introduction of the VeloRouter in major cities in the Baltic Sea region. The choice of a large region will allow to test the effectiveness of the software and reduce potential mistakes as well as bottle-neck situation opportunities. Sustainability assessment methodology IASAM will be disseminated and responsiveness of the tools will be developed to adapt to smart phones capabilities. The authors intend to make the IASAM tool a convenient and successful instrument for various sectors of the economy. The IASAM will allow users, without special background knowledge, to make a macroeconomic assessment of potential sustainability trends scenario projections in real time.

ACKNOWLEDGMENT

The article publication is initiated by FP7 FLAG-ERA FuturICT 2.0 (2017-2020) "Large scale experiments and simulations for the second generation of FuturICT" STIMUL projects "System dynamics simulation use in sociotechnical systems acceptance and sustainability assessment (IASAM3)" and "Intermodal bicycles network as efficient kind of urban green transport (VeloRouter 2.0)".

REFERENCES

[1] United States Environmental Protection Agency, *Global greenhouse gas emissions data 2018*. Available: , https://www.epa.gov/ghgemissions/global-greenhouse-gas-emissions-data. [Accessed: March 18, 2018].

[2] European Commission, *European Commission Decision C 6776 Horizon 2020, Part 11. Smart, green and integrated transport.* Available: http://ec.europa.eu/research/participants/data/ref/h2020/wp/2016_2017/main/h2020-wp1617-transport_en.pdf . [Accessed: March 18, 2018].

[3] S. Sonntagbauer, K. Nazemi, P. Sonntagbauer, and D. Burkhardt, "The FUPOL policy lifecycle," *in Handbook of Research on Advanced ICT Integration for Governance and Policy Modeling*, IGI Global, 2014, pp. 61-86.

[4] M. Sutton, "33 key cities where cycling is growing its modal share," 2016. [Online]. Available: https://cyclingindustry.news/five-key-cities-where-cycling-is-taking-modal-share-from-cars/. [Accessed: March 18, 2018].

[5] E. Ginters, M. Baltruks, I. Sakne, and Y. Merkuryev, "Dual use bicycle path network designing and exploitation environment - VeloRouter," in *Proc. of the European Modeling and Simulation Symposium, 26-28 Sept., Cyprus, 2016*, Bruzzone, Jiménez, Longo, Louca, and Zhang, Eds. ISBN 978-88-97999-68-3, 2016, pp. 10-14.

[6] E. Ginters, Y. Merkuryev, M. Baltruks, and S. Sonntagbauer, "Intelligence enhancing of dual use bicycle routes designing and planning system simulator," *Procedia Computer Science*, vol. 104, pp. 525-529, 2017.

[7] K. Nazemi, M. Breyer, D. Burkhardt, C. Stab, J. Kohlhammer, "SemaVis: A new approach for visualizing semantic information," in *Towards the Internet of Services: The THESEUS Research Program*, Springer, 2014, pp. 191-202.

[8] A. Aizstrauts, E. Ginters, D. Aizstrauta, and P. Sonntagbauer, "Easy Communication Environment on the cloud as distributed simulation infrastructure," *in Proc. of the 5th WSEAS World Congress on Applied Computing Conference (ACC '12), 2-4 May, Faro, Portugal, 2012.* ISBN: 978-1-61804-089-3, ISSN: 1790-5109, 2012. pp. 173-179.

[9] K. Pan, S.J. Turner, W. Cai, and Z. Li, *A Hybrid HLA Time management algorithm based on both conditional and unconditional information,* Parallel & Distrib. Comput. Center, Nanyang Technol. Univ., Singapore, 2008, pp. 203-211.

[10] V. Issarny, N. Georgantas, S. Hachem, A. Zarras, and P. Vassiliadis et al., "Service-oriented middleware for the Future Internet: State of the art and research directions," *Journal of Internet Services and Applications*, 2 (1), Springer, 2011, pp. 23-45.

[11] S.K. Cohen, "Innovation-driven industry life cycles," in *Encyclopedia of technology and innovation management*, V. K. Narayanan, G. Colarelli O'Connor, Eds. Wiltshire: John Wiley&Sons, 2010, pp. 69-76.

[12] M. Taylor and A. Taylor, "The technology life cycle: Conceptualization and managerial implications," *International journal of production economics*, vol. 140, Issue 1, pp. 541–553, 2012.

[13] J. Mac Vaugh, F. Schiavone, "Limits to the diffusion of innovation. A Literature Review and Integrative Model," *European Journal of Innovation Management*, 13(2), pp. 197-221, 2010.

[14] V. Venkatesh, M. Morris, G.B. Davis, and F.D. Davis, "User acceptance of information technology: Toward a unified view," *MIS Quarterly*, 27 (3), pp. 425-478, 2003.

[15] V. Venkatesh, H. Bala, "Technology acceptance model 3 and a research agenda on interventions", *Decision Sciences*, 39 (2), pp. 273–315, 2008.

[16] D. Aizstrauta, E. Ginters, "Sociotechnical systems acceptance and sustainability assessment using dynamic simulation model," in *Proc. of the European Modeling and Simulation Symposium*, Affenzeller, Bruzzone, Jiménez, Longo, and Piera, Eds. ISBN 978-88-97999-85-0, 2017, pp. 519-527.

[17] Z. Barkane, E. Ginters, "Introduction to technology acceptance and sustainability modelling," in *Annual Proceedings of Vidzeme University of Applied Sciences "ICTE in Regional Development"*, 2009/2010, pp.62-90.

[18] S. Fortmann-Roe, "Insight Maker: A general-purpose tool for web-based modeling & simulation," *Simulation Modelling Practice and Theory*, vol. 47, pp. 28-45, 2014.

[19] E. Ginters, D. Aizstrauta, "Technologies sustainability modeling," in *Trends and Advances in Information Systems and Technologies*, vol.2, Rocha, Adeli, Reis, and Costanzo, Eds. ISSN 2194-5357, Springer, 2018, pp. 659-669.

CMedia Compressor: An Application to Graphically Compare General Compression Algorithms and Adaptive Huffman Compression Algorithm

Sampson D. Asare
Department of Computer Science
University of Botswana
Gaborone, Botswana
asaresd@mopipi.ub.bw

Phyela Mbewe
Department of Computer Science,
University of Botswana
Gaborone, Botswana
phy.mbewe@gmail.com

***Abstract*—Much discussion and research has been done on compression algorithms but not many have gone to the extent of developing an application that can test and compare various compression algorithms with each other. In this paper, we present an application known as the CMedia Compressor, that tests and compares various compression algorithms. The application shows that adaptive Huffman and Arithmetic Coding tend to outperform the others such as the Run-length encoding (RLE), Huffman coding, Arithmetic coding (AC) and Lempel-Ziv-Welch (LZW) methods. Results obtained from this study are discussed and also presented graphically. The study's results indicates that Arithmetic coding yields higher space-savings than Huffman coding and Adaptive Huffman coding.**

Keywords—adaptive Huffman, Arithmetic coding, compression, run-length, LZW, algorithm, computer graphics

I. INTRODUCTION

Data compression algorithms re-encode file-data into compact representations of the same data which is intended to reduce the file size for better storage and online transmission of the file [1]. Weizheng et al. [2] deduced that it is necessary to use an appropriate compression algorithm to cut the communication cost of transmitting data over a network. The goal of our project was to develop an application for image compression which compares existing image compression techniques, with a graphical user interface, to determine the effectiveness of different compression algorithms.

This project included the implementation of an application which was used to compare the existing compression methods and to analyze image compression based on time complexity (time taken to complete the compression), reduction in the image file size and the quality of the compressed image. The application was implemented by the researchers to assist with data collection; the application has been titled '*CMedia Compressor*' (*Custom Media Compressor*) and it enables users to choose between a few common general compression methods, thus changing how an input file will be compressed. The general compression methods included are: *Huffman coding*, *Adaptive Huffman coding*, *Arithmetic coding*, *Run-length encoding (RLE*) and *Lempel-Ziv-Welch (LZW)* compression.

The CMedia Compressor displays compression results, via text and a graph; the file size reduction between the input file and the compressed file as well as displaying the time taken to compress the input file by the chosen compression method. The main focus of this project was to analyze and compare the compression of BMP image files. However, the compression of different text files and audio files was implemented into the CMedia Compressor for analyzed and compared. Additionally, to mitigate the effect of background processes giving a false compression-result of the time it takes to compress a file, each compression was carried out at least five times then an average time taken to compress the file was used as the final result of the compression method's execution time.

According to Chuang et al. [3], established lossless general compression methods Run-length encoding (RLE) , Huffman coding, Arithmetic coding (AC) and Lempel-Ziv-Welch (LZW) compression have all contributed towards improving lossless data compression techniques. These methods have been studied in order to gain better understand of data compression and in some cases variants of these methods have been implemented to improve compression. For example, the Adaptive Huffman coding algorithm was implemented to mitigate the flaws of the Huffman coding algorithm. This project mainly focused on comparing the Adaptive Huffman coding algorithm against and in-combination with other general compression algorithms.

II. LITERATURE REVIEW AND RELATED WORK

A. Huffman Coding and Adaptive Huffman Coding Overview

The 'Huffman coding' algorithm (also known as Huffman compression or Huffman encoding) uses the frequency (number of occurrence) of individual characters, in the data, to compress the data [4]. For characters that are more frequent than others, Huffman coding uses fewer bits to encode those characters and then uses more bits to encode the less frequent characters. Compressing characters can also be applied in the compression of BMP image files; all the files stored in memory are stored in bytes, each single byte has a corresponding character and vice versa. (For example the byte '*01000010*' should constantly represent the character 'B' within computer systems using the ASCII international standard).

Adaptive Huffman coding (also known as Adaptive Huffman compression) was established by Faller, Gallager and Knuth. As such, it was also referred to as the FGK algorithm [5]. The Huffman coding algorithm was extended because it required prior knowledge of statistical information (the number of occurrences of each character in the file) and such information is not available for certain files (such as files being streamed over a network) [6]. With the Adaptive Huffman coding algorithm: as characters are processed, frequencies are updated, and character-codes are changed dynamically while the algorithm is being executed. Concerning Adaptive Huffman coding, Gough and Lu [7] presented a faster implementation of Adaptive Huffman coding which utilized two binary trees instead of one. Their experimental results showcase that their implemented algorithm was faster than the general Adaptive Huffman coding algorithm. However, the results do not include execution–time comparisons of the faster Adaptive Huffman coding algorithm against other compression methods, such as Huffman coding, LZW, Arithmetic coding etc.

B. *Arithmetic Coding Overview*

Unlike Huffman coding, 'Arithmetic coding' (AC) does not use a discrete number of bits for representing each symbol to compress data [8]. Arithmetic coding assigns an interval to each symbol, starting with the interval [0...1): zero (inclusive) up to one (exclusive). Each interval is divided into subintervals with sizes proportional to the current probability of the symbols. The subinterval from the coded symbols is then used as the interval for the next symbol; the output will be the interval of the last symbol. The entire data to be encoded is therefore represented by a rational number [9].

Shahbahrami et al. deduced that the Arithmetic coding algorithm produces a more efficient compression (smaller image file sizes) than the Huffman coding algorithm however Huffman coding requires less execution time than Arithmetic coding [10]. Despite the fact that compression from Arithmetic coding usually yields greater reduction in image file sizes than Huffman coding; the majority of JPEG files have used Huffman coding mainly because of the software patents which were associated with Arithmetic coding [11] Arithmetic coding (when used in JPEG files) had multiple software patents which have only recently expired (after which most web-browsers and JPEG viewer-applications have already incorporated support for Huffman coding based JPEG files).

Howard and Vitter [12] deduced that the output from Ziv-Lempel coding can be further compressed by Arithmetic coding. However, the disadvantage of this is that the algorithm is slowed down and has an increased complexity. They developed a fast coder based on a reduced-precision variant of Arithmetic coding with a minimal loss of compression efficiency [12]. Essentially Howard and Vitter implemented faster Arithmetic coding variants, however similar to Gough and Lu's faster Adaptive Huffman coding variant, execution-time comparisons of these faster Arithmetic coding implementations against other compression methods, such as Huffman coding, LZW etc., is not specified

C. *Lempel-Ziv Welch (LZW) Overview*

Many files have strings which are repeated very often, for example the string " *the* " (including the spaces) would take 5 bytes, or 40 bits. If the whole string is given a code, such as 256 (a binary-code of 9 bits: `100000000`), then every time we encounter " *the* " in the data, it could be substituted with the code 256. This code would take 9 bits instead of 40. 'Lempel-Ziv-Welch' (LZW) compression utilizes this concept: it begins with a dictionary of single characters with indexes 0 to 255; it then starts expanding the dictionary such that redundant/repeating strings will be coded, and thereby the original file will be compressed [13].

The algorithm's name comes from its developers: Jacob Ziv and Abraham Lempel developed the dictionary-based compression algorithms LZ77 and LZ78 in 1977 and 1978 respectively [13]. Terry Welch improved LZ78 in 1984 and this algorithm became known as the LZW algorithm. LZW compression is used in GIF image files and is optionally in TIFF and Postscript files [14]. The achievable compression rate depends on the repeating sequences in the data to be compressed [15]

III. RESEARCH METHODOLOGY

A. *The Research Design*

The research design focuses on experiments; a set of image files at different file sizes were sampled and utilized in the experiments. A quantitative research approach was utilized for this study, focusing on numerical analysis of results obtained from experiments. A set of image compression algorithms (also referred to as compression methods within this document) was identified and used in the experiments. The experimentations then utilized these image compression methods in order to identify the image compression methods which best compress the image files, taking into consideration the file size reduction and the time complexity of the algorithms experimented. The *CMedia Compressor* application was used to compared the implementations of the Adaptive Huffman coding algorithm against the static Huffman coding algorithm as well as comparisons with other existing and conventional image compression methods.

As indicated in Figure 1 and Figure 2, the CMedia Compressor application enables users to utilize five general compression algorithms in a simple to use interface and thus allowing users to identity and formulate their own opinions, conclusions and understandings of the existing general compression methods, in addition to simply using the application to compress files of certain file types.

B. *The Sample*

For consistency with other image processing research studies, four images where selected which are commonly used for image processing studies. These images are in the public domain; this means that, similar to images which have a CC0 license, they may be used and distributed for a variety of purposes with minimal restrictions. All sample images are bundled with the CMedia Compressor application for replication of the compression results obtained within this study.

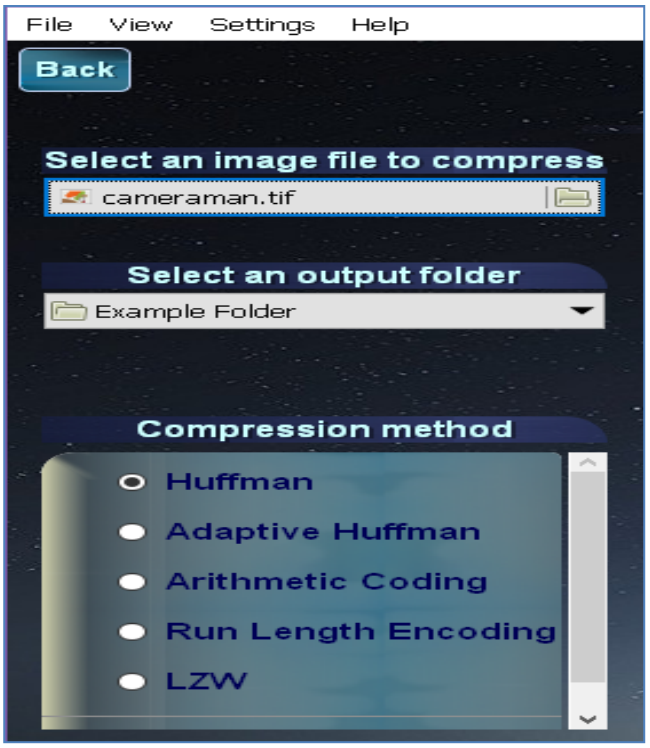


Figure 1. CMedia Compressor screenshot; compression method selection

The sample images used within this study are within the public domain, have a CC0 license or they are owned by this study's researcher; this has been done such that all the images and sample files can be used, re-used or distributed without restrictions. Mostly focusing on the compression, the samples include those as shown in Table I.

C. *Data Collection*

Data was obtained from the experimental results of the compression algorithms implemented into the *CMedia Compressor* application which was implemented by this study's researchers. Statistical data such as file size reduction and the time taken to complete a compression attempt was collected, tabulated and graphed to efficiently compare the image compression algorithms. The second variable, *file size reduction*, was converted into a format known as *'space-savings'*. Space-savings is calculated using a formula:

space-savings = (1 – (compressed-file-size / uncompressed-file-size)) x 100

Space-savings indicated the file size reduction as a percentage. For example, if a compressed file yields a space-saving of 60%, this means that (based on 100% - 60%) the final compressed file had a file size which was 40% of the original uncompressed file size.

IV. GRAPHICAL RESULTS FROM THE EXPERIMENTS

A. *Compression Results Comparison Overview*

The sample image file `cameraman.tif` was compressed multiple times with each of the five selected compression methods and the results of this compression are indicated in Figure 3 and Table 2. For each compression attempt, two variables, compression time and space-savings were recorded and tabulated for the `cameraman.tif` image file. Notably, LZW and RLE compression resulted in negative space-savings values for `cameraman.tif`, this means that the compressed output files were actually larger than the original uncompressed input file, which are bad compression results as they go against the purpose of compressing files altogether.

Secondly the researcher noted that Arithmetic coding yielded the fastest compression time and the highest space-savings. Based on prior knowledge and studies; it was expected that Arithmetic coding would yield higher space-savings than Huffman coding and Adaptive Huffman coding. However Arithmetic coding also yielding the fastest execution time was an unexpected interesting development.

Howard and Vitter [12] indicated that they had implemented a faster Arithmetic coding variant which yielded a faster compression time when compared with a general common implementation of Arithmetic coding. The researchers deduced that the *CMedia Compressor*, used to compress `cameraman.tif`, was using a fast Arithmetic coding variant which yielded a faster compression time than Huffman coding and still withheld Arithmetic coding's ability to produce a higher space-savings value than the other four compression methods. This may have been a once-off occurrence specific to `cameraman.tif`.

In a similar manner to `cameraman.tif`; three other image sample files were compressed multiple times with each of the five selected compression methods and the compression results are indicated in Table 3. For the three other image sample files, the fast Arithmetic coding variant yielded the fastest compression time and the highest space savings result consistently

The researchers compressed the text sample files with each of the five selected compression methods and the compression results are indicated in Table 4. This was done to add some variety to the study and assist in gathering an overall viewpoint of the compression methods' results for different file-types. However, the main focus of the study was on the compression results from image files. The compression results for `text-test [1.0 KB, ASCII].txt` are also displayed graphically in Figure 4.

Unlike the relatively larger image files, the Adaptive Huffman coding algorithm performed more efficiently with the smaller text files; 512 bytes, 1 kB and 10 kB text files. Adaptive Huffman coding yielded much shorter compression time results for smaller files (below 1 MB). The fast Arithmetic coding (AC) variant yielded negative space-savings for the smaller, 512 bytes, text files. While the fast AC variant yielded the best results for image (Megabyte-sized) files, it was not suited to the smaller (below 10 kB) files.

Figure 2. CMedia Compressor screenshot; compression results displayed via graph and text

TABLE I. THE SAMPLE, FILENAMES, FILE TYPES AND LICENSING CONSIDERATIONS

File Category	Filename / File type	Licensing Considerations
Images:		
4 standard common image-processing test images	*cameraman.tif*	Public Domain
	mandril.tif	Public Domain
	peppers.tif	Public Domain
	peppers_gray.tif	Public Domain
Text:		
4 basic ASCII text files (with specific file sizes)	*text-test [1.0 KB, ASCII].txt*	CC0 license / Generated and owned by this study's researchers

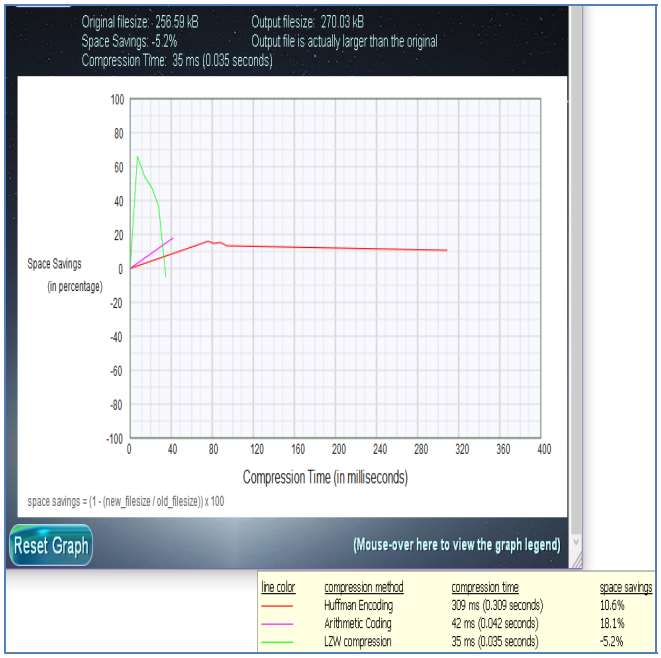


Figure 3. CMedia Compressor screenshot; graphical compression results of Huffman coding, Arithmetic coding and LZW used to compress `cameraman.tif`

TABLE II. COMPRESSION METHODS COMPARISON OVERVIEW FOR *CAMERAMAN.TIF*

Filename	Compression Method	Compression Time (seconds)	Space-savings (%)	Overall Rating
mandril.tif	Huffman	0.740	10.50	100.56
	Adaptive Huffman	4.669	10.70	93.30
	Arithmetic coding	0.072	13.40	110.60
	RLE	0.375	-0.70	67.69
	LZW	0.122	-7.00	49.30
peppers.tif	Huffman	0.456	39.30	278.67
	Adaptive Huffman	2.241	39.60	276.00
	Arithmetic coding	0.056	52.40	318.77
	RLE	0.234	0.00	161.21
	LZW	0.078	18.90	218.23
peppers_gray.tif	Huffman	0.344	44.60	304.49
	Adaptive Huffman	1.691	44.80	302.40
	Arithmetic coding	0.044	56.00	339.29
	RLE	0.181	0.00	171.02
	LZW	0.069	36.30	280.14

TABLE III. COMPRESSION METHODS COMPARISON OVERVIEW FOR MULTI IMAGE SAMPLE FILES

Compression Method	Compression Time (seconds)	Space-savings (file size reduction) %
Huffman coding	0.254	10.60
Adaptive Huffman coding	1.146	11.20
Arithmetic coding (AC)	0.043	18.10
Run-length encoding (RLE)	0.167	-4.90
Lempel-Ziv-Welch (LZW)	0.037	-5.20

TABLE IV. COMPRESSION METHODS COMPARISON OVERVIEW FOR MULTI TEXT SAMPLE FILES

Filename	Compression Method	Compression Time (seconds)	Space-savings (%)	Overall Rating
text-test [1.0 kB, ASCII].txt	Huffman	0.087	-5.50	99.90
	Adaptive Huffman	0.007	33.30	216.46
	Arithmetic coding	0.025	-1.20	112.93
	RLE	0.001	-1.30	112.67
	LZW	0.004	11.20	150.17
text-test [10.0 kB, ASCII].txt	Huffman	0.100	39.00	285.00
	Adaptive Huffman	0.031	42.30	295.04
	Arithmetic coding	0.032	36.40	277.34
	RLE	0.007	-0.90	165.49
	LZW	0.004	55.10	333.49
text-test [20.0 kB, ASCII].txt	Huffman	0.097	40.60	307.80
	Adaptive Huffman	0.072	42.40	313.25
	Arithmetic coding	0.031	43.40	316.33
	RLE	0.010	-0.90	183.48
	LZW	0.004	61.10	369.49
text-test [512 bytes, ASCII].txt	Huffman	0.090	-32.40	87.41
	Adaptive Huffman	0.001	29.10	272.08
	Arithmetic coding	0.028	-3.90	173.03
	RLE	0.001	-1.00	181.78
	LZW	0.143	-0.60	182.70

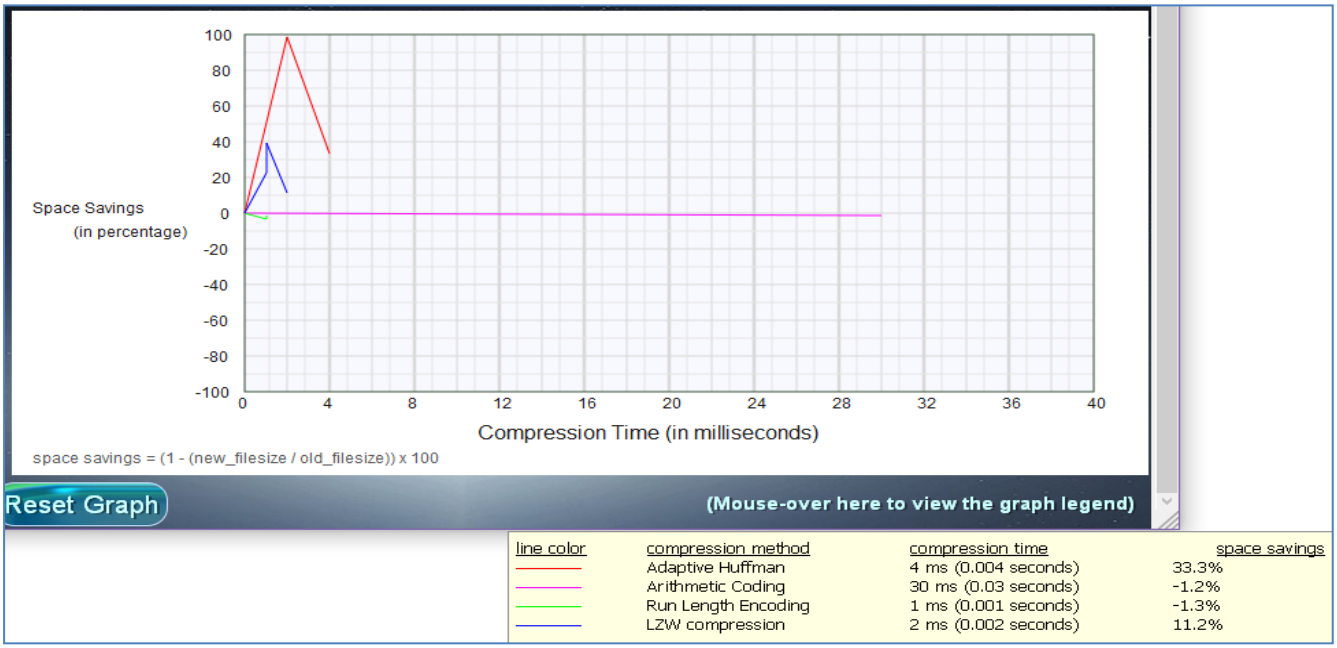


Figure 4. CMedia Compressor Screenshot; Text-Test]1.0 KB, ASCII].txt Graphical Compression Results

V. CONCLUSION

The study's results indicated that Arithmetic coding yields higher space-savings than Huffman coding and Adaptive Huffman coding. These results matched with prior studies as [10] observed that Arithmetic coding will usually provide a better compression result than Huffman coding. Incidentally, the study also matches Howard and Vitter's observation [12] that there are faster Arithmetic coding variants which mitigate the biggest drawback of Arithmetic coding. Howard and Vitter noted that the disadvantage of Arithmetic coding is its slow speed and complexity. This study's findings indicate that a fast AC variant yielded both faster compression times and higher space-savings than Adaptive Huffman coding and Huffman coding, for most of the sampled image files. Fast AC's overall efficiency was lower for text files than for image files, as it was also observed that it no longer yielded good compression results for smaller (less than 10 kB) text files. This means that while fast AC may be the most suitable general compression method for uncompressed image files; it is not the most suitable compression method to be used for every file-type nor every file, especially not for smaller files.

VI. FUTURE WORK AND ACKNOWLEDGEMENT

This study mainly focused on comparing existing general compression methods for image files. In the future, the development of a new algorithm, which compresses video files, can be explored, with the intention of contributing towards the improvement of video compression. This conceptual algorithm would be derived from or inspired by a fast Arithmetic coding variant, the analysis of other existing video compression methods and a dynamic variant of vector quantization (whereby the effect of vector quantization could be adjusted to only affect certain regions or certain frames of the video).

REFERENCES

[1] *Research-Article-Image-Compression* (2014, July) [Online] https://www.omicsonline.org/telecommunication/research-articles-image-compression.php

[2] R. Weizheng, W. Haobo., X. Lianming and C. Yansong, "*Research on a Quasi-Lossless Compression Algorithm Based on Huffman Coding", in* Proc. of International Conference on Transportation, Mechanical, and Electrical Engineering (TMEE), Changchun, China, 16-18 Dec, 2011.

[3] C.P. Chuang, G. Chen., Y. Liao and C. Lin, "*A Lossless Color Image Compression Algorithm with Adaptive Arithmetic Coding Based on Adjacent Data Probability",* in Proc. of International Symposium on Computer, Consumer and Control, Taichung, Taiwan, 4-6 June 2012.

[4] A. Allain. (2011) *Cprogramming.com*. [Online]. http://www.cprogramming.com/tutorial/computersciencetheory/huffman.html

[5] G.H. Ong and W.T. Chong, "*Compressing chinese text files using an adaptive huffman coding scheme and a static dictionary of character pairs",* in International Conf. on Information Engineering, 1993.

[6] D. Marshal. (2001, April) www.cs.cf.ac.uk. [Online]. http://www.cs.cf.ac.uk/Dave/Multimedia/node212.html

[7] W.W. Lu and M.P. Gough, *A Fast-Adaptive Huffman Coding Algorithm,* IEEE Trans. on Communications, vol. 41, no. 4, 1993.

[8] *Arithmetic Coding,* www.data-compression.info.. [Online]. http://www.data-compression.info/Algorithms/AC/

[9] P. Ezhilarasu, N. Krishnaraj and B. Dhiyamesh, "*Arithmetic Coding for Lossless Data Compression – A Review"*, International Journal of Computer Science Trends and Technology, vol. 3, no. 3, 2015.

[10] A. Shahbahrami, R. Bahrampour., M. S. Rostami, and M. A. Mobarhan , "*Evaluation of Huffman and Arithmetic Algorithms for Multimedia Compression Standards,"* https://arxiv.org/ftp/arxiv/papers/1109/1109.0216.pdf.

[11] M. Sivaraman. (2010, October). [Online]. https://lists.fedoraproject.org/pipermail/devel/2010-Oct/143688.html

[12] P.G. Howard and J.S. Vitter, "*Practical Implementation of Arithmetic Coding",* The Kluwer International Series in Engineering and Computer Science, vol. 176, pp. 85-112, 1992.

[13] *Interactive LZW Compression (2018 August).* [Online]. http://www.cs.sfu.ca/CourseCentral/365/li/squeeze/LZW.html

[14] S. W. Smith. (2011) *The Scientist and Engineer's Guide to Digital Signal Processing,* California Technical Publishing. [Online]. http://www.dspguide.com/ch27/5.htm

[15] P. Ezhilarasu and P. K. Kumar, "*LZW Lossless Text Data Compression Algorithm – A Review",* International Journal of Computer Science & Engineering, vol. 6, no. 11, Nov 2015.

The Factors Influencing Customers to Conduct Online Shopping: South African Perspective

Fhatuwani Vivian Mapande and Martin Appiah
Information and Communication Technology
Vaal University of Technology
Vanderbijl Park, South Africa
vivica.mapande664@gmail.com, martina@vut.ac.za

***Abstract*—E-commerce has many advantages reflected in our life, as it makes the daily life of people convenient. This paper aims to investigate the factors influencing customers to conduct online shopping. The paper used the quantitative approach where questionnaires were used to gather information in the format of a survey. For e-commerce adoption factors, Diffusion of Innovation (DOI) model was used to assess participants' perspectives when it comes to the adoption of e-commerce. More specifically, the paper proposed the framework for South African customers. The proposed framework was tested on a sample of South African residents. The final sample consisted of 235 respondents. The reliability of the questionnaire was thoroughly examined. Empirical data were analyzed using the linear regression. The results of the research revealed that privacy & security, customer trust, perceived service quality, IT knowledge and Relative advantage are significant factors in influencing customers to adopt e-commerce. On the other hand, privacy was found not to be the statistically significant impact on the adoption of e-commerce**.

Keywords—Security, Adoption, Diffusion of Innovation (DOI), E-commerce

I. INTRODUCTION

E-commerce has numerous benefits reflected in our life, as it makes the day by day life of individuals advantageous. E-commerce alludes to the way toward purchasing or offering items over the Internet [3]. Web based shopping is ending up progressively prominent as a result of speed and usability for customers. [3]. It empowers the purchasing and offering of items through the retail storefronts. It is tied in with utilizing the convenience, accessibility and worldwide reaches to enhance existing organizations or to make the new virtual business. Moreover, E-commerce is in a general sense an extra sales channel, the key highlights of which are speed, adaptability, and transparency [4].

As a rule, protection, security, and trust are most critical feelings of fear for online technologies [8]. Most organizations apply E-commerce security to the components that influence E-commerce to be correct; PC security, information security, honesty, accessibility and other Information Security framework. The Business to Consumer (B2C) Internet commerce appreciates an unfaltering development rate (around 19% every year), and it is a natural method of shopping for some shoppers [40].

E-commerce provides greater opportunities to businesses such as greater visibility, broader market reach, reduced warehousing costs, shorter delivery times, more efficient procurement processes and higher customer satisfaction [5]. The demand of various online services is rapidly increasing, this trend is changing the way business is conducted between businesses and individuals [1]. The adoption of online services by businesses and individuals is becoming the need in the current digital world, this is supported in the study conducted by [4].

Customers are concern with security, privacy, safety and trust issues that comes with online transactions, whereby, businesses has to implement security measures to protect customers online [6]. Besides, Transactions on the E-commerce are constantly conducted through the web for the reason that the business ought to have a top of the line security framework and ought to have a good protection control framework, keeping in mind the end goal to ensure trust between its clients [6].

A. Research problem

The researcher has found through different investigations that E-commerce has distinctive perspectives with regards to its security. Besides, the faltering of privacy loss assume a vital part amid the purchasing process [7, 8]. Individuals who feel security is an issue in internet shopping don't confide in the E-commerce [7]. Likewise, a great many people don't see the significance of e-commerce. Security influences people on the regular basis, whereby people including clients, and organizations adjusting the new technology advancements will prompt shopping success [7]. Be that as it may, organizations should address distinctive new security challenges and be sure to keep up the most noteworthy standards of e-commerce security, to protect both themselves and their clients [6].

B. Objectives

This paper consists of the following objectives:

- To investigate various technology adoption models.
- To propose the research model for online shopping customers in South Africa.
- To evaluate the proposed model and determine factors influencing customers to conduct online shopping in South Africa.

II. Literature Review

A. E-commerce in South Africa

E-commerce in South Africa is developing quickly with R10bn clients amid 2017. The advanced digital commerce development opportunity is exceeding expectations in business industry. High mobile penetration is driving the increase of buyer trust in online transactions [9], this is driving buyers' shopping behaviour change as well. Moreover, Local web retailers are changing channels to have the capacity to increase the value of their buyer base and they are rearranging and enhancing availability and offering more advantages [9]. Though, organizations are enthusiastically executing an assorted variety of computerized activities to change the client experience [9].

South African online spend is determined to develop to more than 53 billion Rand by 2018 with an anticipated yearly development rate of 15% through 2021 [4]. This records for 1% of the nation's overall retail division and shows the colossal potential for development and opportunity. Besides, Internet client penetration is at 47% and anticipated that would achieve 60% by 2021 and mobile entrance is 65% and developing [10]. In South Africa, Business to Business (B2B), Business to Customers (B2C) relations is by and large the same. Both have gotten familiar to performing buyer product research on the web. Likewise, the two kinds of clients are utilizing consumer and B2B sites to purchase products and services for their organizations or as people [10].

B. Technology Adoption Models

With regards to technological development and innovation adoption, there is a developing number of hypotheses and studies. The most common theories are Technology Acceptance Model 3 (TAM3)[26], Diffusion of Innovation Theory (DOI) [19], Theory of Reasoned Action (TRA)[20], Technology Acceptance Model (TAM)[24], Technology Acceptance Model 2 (TAM2)[25], Theory of Planned Behavior (TPB) [23], Technology-Organization-Environment (TOE) Model [22] and The Unified Theory of Acceptance and Use of Technology (UTAUT) display [21]. These models are utilized to inspect adoption and user acceptance of technology innovation [41, 42]. Also, numerous researchers perceive DOI in analyzing the user behaviour of innovation adoption.

In this paper we considered Diffusion of Innovation model to determine factors influencing customers to conduct online shopping in South Africa. DOI model is the appropriate model to products that hold possible usage in technology applications [28]. Various industries such as banking and e-business used DOI for predicting the level of adoption [35, 36]. Innovation itself, communication networks, time necessary are four core fundamentals that control the spread of an innovation [17]. In addition, there is the necessary motivation of knowledge, habits, social and personality attributes [17]. Then, the endurance is not ensured. Moreover, there are five main factors of DOI affecting the channel of rising technologies from the macro level, which are the Relative Advantage, Compatibility, Complexity, Observability and Trialability [18].

C. Proposed Research Model

After evaluating the literature on technology adoption and other previous studies the succeeding research model was proposed, derived from DOI. See Fig. 1 below:

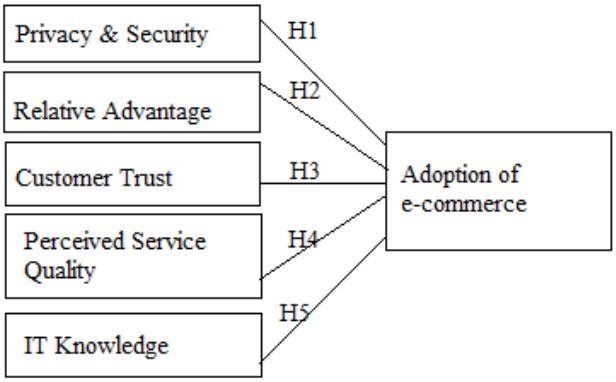


Fig. 1. *Proposed Model*

D. Privacy and Security(PS)

There are numerous styles for fulfilling the privacy and security in the E-commerce, whereby people and shopper groups are calling for new privacy shields for the Internet and other PC systems [43]. Additionally, Personal data that might bear some significance with organizations or individuals with pernicious points is created at whatever point individuals get to the Internet. E-commerce frameworks give security mechanisms, this will diminish the danger of client related data theft which may prompt fraud [14]. E-commerce security is the assurance of web based business resources from unapproved access, utilize, modification, or destruction [6].

Poor security on E-commerce web servers and in clients' personal mobile devices are the centre issues to determine for growth of E-commerce [13]. Therefore, many people are scared to reveal their personal information based on an assessment of risk. Thus, it refers to the extent to which consumers believes that electronic commerce will not compromise his or her privacy [13]. This can be hypothesized as follows:

H1: The adoption of E-commerce is significantly influenced by the customer's privacy and security.

E. Relative Advantage(RA)

Relative advantage is characterized as how much a technology advancement is seen as better or better than ever previously. This can be estimated from any fragment, for example, financial, social distinction, solace, and fulfillment. The relative advantage seen by the adopters, the speedier these technology advancements can be adopted [34]. Besides, with this it tends to be presumed that the relative advantage that is fundamentally connected with attitudes [34]. H2: The adoption of E-commerce is significantly influenced by the customer's Relative Advantage

F. Customer Trust(CT)

Client Trust alludes to an arrangement of convictions about the kindness, capacity and honesty of another party [44]. Client Trust progresses after some time as the client perceives that the supplier is dependable and performs in an upright and legit way [44]; it is a basic factor for making fulfilled and faithful clients in e-commerce [44]. Hence, Adequate encounters with suppliers make innumerable levels of trust, which impacts long term relationships [44]. Trust is progressively vital as online clients live in a hazardous situation of security intrusion and fraud. E-commerce can be fruitful just if PC produced conditions gain the trust of general society [12]. Individual information can be gathered from open profiles that clients post about themselves on social networking sites [11]. Trust assumes an essential part in numerous social and economic communications including vulnerability and reliance [7]. In this study we tested the following hypothesis:

H3: The adoption of E-commerce is significantly influenced by the customer's trust.

G. Perceived Service Quality (PSQ)

E-commerce service quality is the assessments from the genuine experience of the administration as far as excellence and quality of the E-commerce benefit conveyance in the online market [33, 45]. It might be part into specialized quality and practical quality [33]. Service quality has been viewed as a noteworthy factor of achievement for the organizations through which the organizations can detail their upper hand, and uplift their competitiveness [33]. Furthermore, the correct navigational attributes majorly affect client attributes towards online business websites [45].These concepts lead to formulating the following hypothesis:

H4: The adoption of E-commerce is significantly influenced by the Perceived Service Quality.

H. IT Knowledge(ITK)

Innovation can be tended to through learning, that is of new thoughts, new innovation, the new way and new practices that guide and make development [27]. Information is connected to the idea of innovation and can be regarded by the know-how and furthermore verifiable learning for the act of something [27]. These has led to the formulation of the following hypothesis:

H5: The adoption of E-commerce is significantly influenced by the customers' IT knowledge

III. Research Methodology

This paper is a quantitative study conducted using questionnaires. Quantitative research is substantial and countable in nature and the designs are predetermined and structured, remaining consistent throughout the study [38]. The population of this study is South African residents. Population refers to a number of people or units from which research information will be obtained [39]. The sample size for the study is 250. Furthermore, convenient sampling is used to select the participants of the study.

Convenience sampling is characterized as a kind of non-irregular sampling where individuals from the objective populace that meet certain practical criteria [37]. The survey is utilized as an exploration instrument for this paper which included statistic data and adopted related questions for online business. The survey can be characterized as the accumulation of data from a sample of people through their responses to questions. The adoption measures and factors in the section B of the survey questionnaire were gotten from past studies on online business adoption studies [29, 30, 31, 32]. The questionnaire of the research was gotten and adjusted from the DOI model, in which the research model is picked in view of its elements, for example, Relative Advantage, IT Knowledge and Security.

IV. Results

A. Demographic information

Out of 250 sample size, a total of 235 individuals returned valid surveys. The above sample size is selected to reduce the time for conducting the study, although the demographics information is needed for generalization, as South Africa is a diversity country which accommodate people from various ethnic groups. Table I below, shows that the participated sample size is nearly gender-balanced, 119 (50.6%) of the participants are female, hence 116 (49.4%) are male. Majority of participants are African (Black) with 228 (97%) of the sample. Participants' age group ranged from 18 to over 46, whereby 123 (52.3%) of participants were between 26-35 of age. In addition, 222 (94.5%) of the participants use the internet more often. Only 98 (41.7%) of the participants hold university qualifications.

TABLE I. Participants Demographic information

Demographic summary			
Category	***Item***	***Frequency***	***Percentage (%)***
Gender	Male	116	49.4
	Female	119	50.6
Age group	18-25	63	26.8
	26-35	123	52.3
	36-45	34	14.5
	Over 45	15	6.4
Ethnic group	African	228	97
	White	5	2.1
	Asian	0	0
	Indian	2	0.9
Educational Level	High school	23	9.8
	Diploma	98	41.7
	Degree	67	28.5
	Other	47	20
Internet Access	Once a week	10	4.3
	More often	222	94.5
	Once a month	3	1.2

B. Reliability test results

After using Cronbach's alpha to measure the internal consistency between all the six study variables, Table II below shows that all the variables are very reliable with an alpha value ranging from 0.753 to 0.921 based on [46].Hence, a very high Cronbach's alpha indicates high correlations among the items in the scale and support the rule of thumb for the reliability test that is 0.70 or higher represents very reliable and consistent [46].

TABLE II. RELIABILITY DETAILS

Reliability test results		
Variable	***Item***	***Cronbach's Alpha***
RA	4	0.921
PS	4	0.870
ITK	4	0.823
CT	4	0.762
PSQ	4	0.884
AC	4	0.753

C. *Hypothesis results*

A linear regression was utilized to survey the quality of the connection between related variables through the importance of path coefficient (β) and significant level (p-value), the dependent variable (Adoption of e-commerce) among five independent variables (Relative Advantage, Customer Trust, IT Knowledge, Privacy and Security, Perceived Service Quality). The results presented in Table III indicate that the results are in support with the previous studies conducted in predicting the adoption of e-commerce such as privacy & security, relative advantage, customer trust, perceived service quality and IT Knowledge whereby both variables positively influence the adoption of e-commerce and are supported with p-value less than 0.005 ($p<0.005$) [2, 8, 15, 16].

Furthermore, these results indicate that construct Customer Trust (CT) with the higher path coefficient (β) of .623 strongly influence the adoption of e-commerce. Likewise, Privacy and Security (PS) with path coefficient (β) of .350 also influence the adoption of e-commerce, followed by IT Knowledge (ITK) with path coefficient (β) of .187, Perceived Service Quality (PSQ) with path coefficient (β) of .177 which influence the adoption of e-commerce and Relative Advantage (RA) least influence the adoption of e-commerce with path coefficient (β) of .136.

TABLE III. HYPOTHESIS TEST RESULTS

Hypothesis test summary						
Dependent variable	***Independent variable***	***Hypothesis***	***p-value(p)***	***t-value(t)***	***Beta-value(β)***	***Results***
AE	PS	H1	.0001	4.173	.350	Accepted
	RA	H2	.0001	10.231	.136	Accepted
	CT	H3	.0001	1.601	.623	Accepted
	PSQ	H4	.0021	2.270	.177	Accepted
	ITK	H5	.0032	2.224	.187	Accepted

V. CONCLUSION

For E-commerce to grow, Internet is one of the causes, although it causes uncertainty to customers. Whereas, customers are worried about trust, privacy and security. This study has applied the research model derived from DOI in accomplish the study objectives. In addition, Trust is alleged to be the crucial component for consumers to adopt and conduct online. The findings of this study indicates that customers are likely to purchase from the internet if they see a higher level of trust, protection and security in online business. Moreover, individuals with a larger amount of IT Knowledge will probably partake in internet business since training would not be required. These findings are in support of the previous findings on e-commerce and provided lights on factors that influence customers to adopt e-commerce [4, 7, 8, 17, 28]. Furthermore, Customer Trust, Perceived Privacy and Security go hand in hand, whereby Customer Trust on the internet is characterised by the Privacy and Security perceived by customers concerning the control of their personal information. Also, websites should display visible security mechanisms to encourage customers to use the website.

VI. RECOMMENDATIONS

E-commerce has to improve trusting beliefs privacy and security protection, as well as data reliability to reach the customer's relaxation. For successful adoption of E-commerce, online businesses should understand the privacy and security, customer trust, the level of users' knowledge when it comes to IT and Relative Advantage that influence the adoption of e-commerce. Likewise, online businesses must ensure that all customers from various age groups, with various educational level and different gender get to adopt e-commerce. Therefore, in order for online businesses to provide the trust and secure setting, they should apply security measures in full force. Also, future research should focus on the development of various e-commerce adoption models. Moreover, comparative research in various demographics would create more noteworthy outcomes or results since this study was only focused on the factors influencing customers to conduct online shopping. Further research could be conducted in order to investigate if demographics has influence on the adoption of e-commerce.

ACKNOWLEDGMENT

I am grateful to Dr. Martin Appiah for his support and encouragement throughout the work.

REFERENCES

[1] R. Shahjee, "The impact of electronic commerce on the business organization," Scholarly Research Journal for interdisciplinary studies, vol.4,no. 27, pp.3130-3140, Nov. 2016.

[2] A. K. Garg and T. Choeu, "The adoption of electronic commerce by small and medium enterprises in Pretoria east", EJISDC, vol.68, no.7, pp. 1-23, 2015.

[3] A. Gupta, "E-commerce : role of e-commerce in today's business," International Journal of Computing and Corporate Research, vol.4, no.1, pp.1-9, Jan. 2014.

[4] L.A. Hussein and A.S.Baharudin, "Factors affecting small and medium enterprises (SMEs) continuance intention to adopt e-commerce in Jordan," International Journal of Advanced and Applied Sciences,vol.4, no. 4, pp.110-117, Mar. 2017.

[5] S.M. Sabri, R. Sulaiman, A. Ahmad and A.Y. Tang, "A comparative study on it outsourcing models for Malaysian SMEs e-business transformation," ARPN Journal of Engineering and Applied Sciences, vol.10, no, 23, pp.17863-17870, 2015.

[6] M. I. Al Ladan, "E-commerce security challenges: A Taxonomy," Journal of Economics, Business and Management, vol. 4, no. 10, pp.589-593, Oct. 2016.

[7] M. Pilík and E. Juřičková, "Trust and security as significant factors influencing online buying behavior in the Czech Republic," Economics management innovation, vol. 8, no.2, pp.31-44, 2016.

[8] P. Gupta and Dr. A. Dubey, "E-commerce- study of privacy, trust, and security from consumer's perspective," International Journal of Computer Science and Mobile Computing, vol. 5, no, 6, pp.224 – 232, Jun. 2016.

[9] C. Smith, "How e-commerce is exploding in SA," Fin24, Mar 2018. [Online]. Available https://www.fin24.com/Economy/how-ecommerce-is-exploding-in-sa-20180316. Accessed: Jul. 18, 2018.

[10] International Trade Administration (ITA), "South Africa Ecommerce," Export.gov,Jul. 19, 2017. [Online]. Available: https://www.export.gov/article?id=South-Africa-ecommerce. Accessed: Jul. 18, 2018.

[11] D. Chaffey and P.R. Smith, Emarketing excellence: planning and optimizing your digital marketing, 4th ed. London: Routledge, 2013.

[12] S. Birzhandi, N. Pour Moallem, and S.J. Ghoreishi, "Evaluation of e-trust building structures interact with transportation," In Proceedings of 2014 8th International Conference on e-Commerce in Developing Countries: With Focus on e-Trust, April 24-25, 2014, Mashhad, Iran, IEEE, 2014, pp-1-8.

[13] J. Singh, " Review of e-commerce security challenges," International Journal of Innovative Research in Computer and Communication Engineering,vol. 2, no.2, pp.2850-2858, Feb. 2014.

[14] G. Gautam and D. Yadav, "Sentiment analysis of twitter data using machine learning approaches and semantic analysis," 7th international conference on Contemporary Computing (IC3), vol. IEEE Xplore, pp.437-442, Sep. 2014.

[15] Y. Choi and D. Q. Mai, "The sustainable role of the e-trust in B2C e-commerce of Vietnam," Sustainability,vol.10, no.1, pp.1-18, Jan. 2018.

[16] L. Chen and R. Wang, "Trust development and transfer from electronic commerce to social commerce: an empirical investigation," American Journal of Industrial and Business Management, vol. 6, no.5, pp. 568-576, May. 2016.

[17] S. Buc and B. Divjak, "Environmental factors in the diffusion of innovation model: e-learning in a higher education institution," Central European conference on information and intelligent systems, September 2016, Varazdin, Croatia, pp.99-250.

[18] S. K. Mathur and H.V. Verma, "Significance of DOI model for adoption of cloud computing", International journal of emerging trends and technology in computer science, 2014, vol. 3, pp.228-232.

[19] E. M. Rogers. Diffusion of Innovations. 3rd ed. New York: The Free Press, 1983.

[20] M. Fishbein and I. Ajzen, Belief, attitude, intention, and behaviour: an introduction to theory and research. MA: Addison-Wesley, 1975.

[21] V. Venkatesh, M. G. Morris, G.B. Davis and F.D. Davis, "User acceptance of information technology: toward a unified view," MIS Quarterly, vol.27, pp.425-478, 2003.

[22] L.G. Tornatzky, M. Fleischer and A. Chakrabarti, A., Processes of technological innovation (Issues in organization and management series). Lexington Books.

[23] I. Ajzen, "Consumer attitudes and behaviour: the theory of planned behavior applied to food consumption decisions," Rivista di Economia Agraria, Anno LXX, no.2, pp. 121-138, 2015.

[24] F.D. Davis, "Perceived usefulness, perceived ease of use, acceptance of information technology," MIS Quarterly, vol.13, no.3, pp. 319-340. Sep.1989.

[25] V.Venkatesh and F.D. Davis, "A theoretical extension of the technology acceptance model: four longitudinal field studies," Management Science, vol.46, no. 2, pp. 186-204, Feb. 2000.

[26] V.Venkatesh and H. Bala, "Technology acceptance model 3 and a research agenda on interventions", Decision Sciences, vol.39, no.2 pp. 273-315, May. 2008.

[27] M. Q. Miranda, J. S. Farias, C. de-A. Schwartz, J.P.L. de Almeida, "Technology adoption in the diffusion of innovations perspective: introduction of an ERP system in a non-profit organization," RAI Revista de Administracao e Inovacao, vol.13, no.1, pp. 48-57, Jan. 2016.

[28] S. Raeisi and M. Lingjie, "Factors influencing m-commerce adoption in China," The international journal of business & management, vol.4, pp.372-384, Mar. 2016.

[29] M. A. Masarweh, L. Al-Qaisi, S. Al-Masaeed, and Z. Hunaiti. "E-commerce adoption at customer level in Jordan: an empirical study of Philadelphia general supplies," (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 7, no. 11, pp. 196-205, 2016.

[30] M.S. Kwadwo, A. T. Martinson, T. T. Evans and A. Esther, "Barriers to e-commerce adoption and implementation strategy: empirical review of small and medium-sized enterprises in Ghana," British Journal of Economics, Management & Trade, vol.13, no.1, pp.1-1, Mar. 2016.

[31] M. S. Kabugumila, S. Lushakuzi, and J. E. Mtui, "E-commerce: an overview of adoption and its effective implementation," International Journal of Business and Social Science,vol. 7, no.4, pp-243-252, April. 2016.

[32] R. Rahayu and J. Day, "E-commerce adoption by SMEs in developing countries: evidence from Indonesia," Eurasian Bus Rev, vol. 7, no.1, pp.25–41, Mar.2017.

[33] S. Moghavvemi, S. T. Lee, S. P. Lee, "Perceived overall service quality and customer satisfaction: a comparative analysis between local and foreign banks in Malaysia," International Journal of Bank Marketing, vol. 36, no, 3, pp.908-930, Jul. 2018.

[34] M. Yunus, "Diffusion of innovation, customer attitudes and intention to use mobile banking," Information and Knowledge Management, vol.4,no. 10, pp.12-18, 2014.

[35] M. Dash, P.B.Bhusan, and S.Samal, "Determinants of customers adoption of mobile banking: an empirical study by integrating diffusion of innovation with attitude", Journal of Internet Banking and commerce, vol.19, no. 3, pp.1-21, Dec. 2014.

[36] B. Ali, N. Baluch, and Z. M. Udin, "The moderating effect of religiosity on the relationship between technology readiness and Diffusion of electronic commerce", Modern applied science, vol.9, no, 12, pp.52-60, Oct. 2015.

[37] C. Kivunja, "Innovative methodologies for 21st-century learning, teaching, and assessment: A convenience sampling investigation into the use of social media technologies in higher education," International Journal of Higher Education, vol. 4, no.2, pp.1-26, Jan. 2015.

[38] D. K. Padgett. Qualitative methods in social work research. Sage Publications, 2016.

[39] S.K. Antwi and K. Hamza, "Qualitative and quantitative research paradigms in business research: A philosophical reflection," European Journal of Business and Management, vol.7, no.3, pp.217-225, 2015.

[40] W. Rachbini, "The impact of consumer trust, perceived risk, perceived benefit on purchase intention and purchase decision," International Journal of Advanced Research (IJAR), vol. 6,no.1, pp.1036-104, Jan. 2018.

[41] H. Taherdoost, "A review of technology acceptance and adoption models and theories," Procedia Manufacturing, vol. 22, pp. 960-967, Oct. 2018.

[42] P.C. Lai, "The literature review of technology adoption models and theories for the novelty technology," Journal of Information Systems and Technology Management, vol.14, no.1, pp. 21-38, April. 2017.

[43] M. Ghayoumi, "Review of security and privacy issues in e-commerce," In Proceedings of the International Conference on e-Learning, e-Business, Enterprise Information Systems, and e-Government (EEE), 2016. The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), pp.156-160.

[44] J. Chaparro-Peláez, Á. Hernández-García and A. Urueña-López, "The role of emotions and trust in service recovery in business-to-consumer electronic commerce," Journal of theoretical and applied electronic commerce research, vol.10, no.2, pp.77-90, 2015.

[45] M. Sobihah, M. Mohamad, N. A. M. Ali and W. Z. W. Ismail, "E-commerce service quality on customer satisfaction, belief and loyalty: a proposal," Mediterranean Journal of Social Sciences, vol.6, no.2, pp.260-266, Mar. 2015.

[46] H. Alwan and A. al-zu'bi, "Determinants of internet banking adoption among customers of commercial banks: an empirical study in the Jordanian banking sector," International Journal of business management., vol.11,no.3, pp.95-104, Feb. 2016.

Criminal Profiling in Digital Forensics: Assumptions, Challenges and Probable Solution

Adedayo M. Balogun
Dept. of Information & Communication Technology
Vaal University of Technology
Vanderbijlpark, South Africa

Tranos Zuva
Dept. of Information & Communication Technology
Vaal University of Technology
Vanderbijlpark, South Africa.

***Abstract*—Cybercrime has been regarded understandably as a consequent compromise that follows the advent and perceived success of the computer and internet technologies. Equally effecting the privacy, trust, finance and welfare of the wealthy and low-income individuals and organizations, this menace has shown no indication of slowing down. Reports across the world have consistently shown exponential increase in the numbers and costs of cyber-incidents, and more worriedly low conviction rates of cybercriminals, over the years. Stakeholders increasingly explore ways to keep up with containing cyber-incidents by devising tools and techniques to increase the overall efficiency of investigations, but the gap keeps getting wider. However, criminal profiling – an investigative technique that has been proven to provide accurate and valuable directions to traditional crime investigations – has not seen a widespread application, including a formal methodology, to cybercrime investigations due to difficulties in its seamless transference. This paper, in a bid to address this problem, seeks to preliminarily identify the exact benefits criminal profiling has brought to successful traditional crime investigations and the benefits it can translate to cybercrime investigations, identify the challenges posed by the cyber-scene to its implementation in cybercrime investigations, and proffer a practicable solution.**

Keywords—cybercrime, investigation, criminal profiling, digital forensics, cyber scene

I. INTRODUCTION

Cybercrime, which is very consistently used interchangeably with computer crime and e-crime, includes any activity whose objective serves an illegal interest, and is perpetrated either by directly or indirectly employing or attacking any device capable of data processing, storage, and/or communication operation, or its contents [1][2][3][4][5][6]. Such activities include hacking, money laundering/fraud, intellectual property theft, child exploitation/pornography, identity theft, cyber stalking/bullying, software/movie piracy, denial of service, and may or may not culminate in cyberterrorism.

However, newer methods like hardware trojans and blockchain mining are emerging for achieving the same cybercriminal objectives as the older known methods. The cybercrime menace continues to constitute a much bigger harm than anticipated when researcher [7] highlighted the potential detriments that the computer and internet technologies would together pose. Computer crimes have become a pandemic razing through the privacy, trust, finance, and welfare of the wealthy and low-income individuals as well as renowned and start-up organizations and governments across the world today.

Reports across the continents have shown a thumping growth rate of cyber-incidents, with costs climbing up from seven hundred and eighty-one million dollars ($781 million) in 2013 [8] to over four hundred billion dollars ($400 billion) in 2016 [9]. More shocking reports have predicted an exponential cybercrime cost of over two trillion dollars ($2 trillion) in 2019 [10][11].

The year 2016, in incident numbers, has seen a combined four hundred and seventy-seven million (477 million) data breach by hacking reported for Linkedin and MySpace in May [12], one hundred and twelve million (112 million) for Badoo in June [13], about five hundred million (500 million) for Yahoo in September [14], fifty five million (55 million) for Philippines Commission of Elections in March [15], and ninety eight thousand (98000) for the South African eThekwini in September [16]. The trend has been similar for most of the other cybercrimes. It becomes daunting when the biggest firms with the most secure defences money can buy still end up being victims of cybercrimes [9], as seen with Google, Yahoo and Linkedin in 2016 [12], leaving the question – Who then is safe from being attacked?

Consequently, digital forensics has emerged to counteract the pervasiveness of cybercrimes. Digital forensics is a heterogeneous conflation of the scientific, engineering, artistic, and legal circles with difficult standardizations due to its rapidly growing and incessantly volatile nature [17][18]. It employs a proven protocol of techniques and tools – such as criminal profiling – to ascertain and report the 'what', 'when', 'whom', 'why' and 'how' of a cyber-incident [19]. Criminal profiling traditionally has arts root in the Roman Empire during the Common Era (CE), before its scientific application gained proponents. The earliest and most popular criminal profiling methodology is the *Malleus Maleficarum* published in 1486. It described the specific characteristics and abilities as well as specific circumstances that may identify criminals. The *Malleus Maleficarum* was basically a criminal profiling methodology for identifying witches, within an overall inquisition to prosecute and punish witches [20]. Criminal profiling uses information about the concerned or other crime scenes to provide a vivid description of whom the perpetrator being is. It describes in detail the demographic, geographic, sociological/behavioural and technical/vocational traits of the perpetrator of a particular crime. The resultant profile could subsequently be applied to identify and connect the perpetrator(s) to a crime being investigated, and has proved to be a success in numerous investigations that incorporated it into their protocols [21][22][23].

The rest of this paper is organized thus: Section 2 presents the facts that motivated this work. Section 3 highlights previous cybercriminal profiling related works. In section 4, objectives and procedures that drive profiling in successfully investigating traditional crimes are discussed in stages. Section 5 presents the issues that make the cyberspace a difficult environment to apply criminal profiling. In section 6, the current solutions to the highlighted issues are shown and a more robust solution is proposed. Recommendations on future work are highlighted in conclusion.

II. Motivation

Cybercrime numbers are actually much more numerous than reports show as disclosure rate by victims is reported to be very low, sometimes unintentionally due to being undetected and other times intentionally in order to avoid further losses associated with organizational downtime, public scrutiny and cost of resources directed at investigation [8][24][25]. Unfortunately, conviction rate for cybercriminals is derisively low due to obsolete laws, cross-jurisdictional complexities, deficient investigator-to-criminal ratio, and long investigation durations that span an average fifteen (15) months [25]. These factors constitute technical and legal complexities that lead to case backlogs and eventual disposals [26], when volatile and time-sensitive evidence has been destroyed due to delays between occurrence and prosecution times. The low conviction rates – less than 9% in the US and less than 5% in India (including 90.3% case pendency and disposal) – provide criminals with a sense of invincibility which can be attributed to the increase in cybercrime perpetration [8][26][27].

Hence, there is a drastic need for change in the ways investigations are handled, to reduce the duration of cybercrime investigation so as to increase the conviction of cybercriminals while still maintaining accuracy of suspect identification. Criminal profiling as an accurate and timely investigative technique [28], has reportedly provided valuable investigative directions where the conventional scientific techniques have been non-suggestive during the investigations of traditional crimes [20][22][29][30]. However, as a relatively new focus area [31], there is a difficulty encountered in the transference of this technique seamlessly to the cyber scene. This is due to the complexities regarding the volatility of pseudo-scenes, vastness of technical contexts, and variation in transnational laws and cultures, among others [32], frustrating the emergence of a formal methodology [33].

III. Related Studies

The success of any scientific technique is largely dependent on a number of factors including repeatability [19]. Repeatability is most likely achieved when there is a set procedure like a methodology or framework for applying the particular technique [6][34]. Hence, a lot of works have gone into ensuring the repeatability of methodologies adapted for cyber investigations in order to ensure the accuracy of results and thus, admissibility as evidence in courts. However, their significance have tended chiefly towards academics with little focus on the technical aspects important to the digital investigation industry.

The FBI method of profiling is one of the earliest behavioural scientific criminal profiling system used by a professional body – the Federal Beareau of Investigation [35][36]. The profiling system employs a 5-stage deductive analysis process in building a criminal profile. It includes a framework that uses initial information assessment to classify the criminal. However, this classification framework has received criticisms regards its inaccuracy and indistinctness [22][37]. The insufficiency in the amount of study and sample, as well as the geographical limitation and non-randomness of sample discountenances the validity of the criminal profiling system [38][39]. The system is also restrictively behavioural with no consideration for technicality, thus rendering it inapplicable to cybercrimes.

The behavioural evidence analysis is a formal criminal profiling framework for traditional crimes. It advocates and implements the deductive approach that is seen as objective, rather than the subjective inductive approach [23]. The BEA sets out to monitor, interpret and arrive at a profile whose objective is to provide the court-appointed investigator with a clear direction as to the person that committed a particular crime. It involves 4 main stages which can be applied distinctly in the context of either investigation or trial [20]. The BEA remains a framework that researchers have adapted in studies to prove the benefit of profiling to digital forensic investigation in the absence of a dedicated cybercriminal profiling framework [40][41]. However, its limited applicability to the technicalities of cybercrimes due to the later emergence of cybercrimes needs to be properly addressed.

Owing to the success of the criminal profiling technique in traditional crimes, there have been discussions about its application to crimes committed in the cyberspace. A few proponent researchers have taking some steps beyond argument for its application, and have adapted the traditional criminal profiling methodologies to the investigation of select cybercrime kinds. Some of those studies are highlighted thus.

In [42], the authors described the contents of a typical criminal profile and highlighted in steps how the crime triangle – motive, means and opportunity – can be used to connect a perpetrator to a crime under investigation. While their work was an argument for cybercriminal profiling as an important technique in digital forensic investigations, it did not mention nor demonstrate to practitioners how a profile could be built for any particular kind of cybercrime.

Authors [32] proposed a profiling methodology that combined iteratively a 4-step deductive approach including victimology, motive, offender characteristic identification and evidence examination with an inductive approach. This was adapted from and almost identical to the behavioural evidence analysis framework proposed by [20], except for the supplementary inductive approach incorporated. Reference [32] argued that the offender characteristics could help predict the modus operandi of a cybercriminal that is then compared alongside the other profile contents generated with known cases, so as to identify cybercriminals (possibly serial) with similar modus operandi, in an inductive approach. Unfortunately, this methodology was not subjected to a demonstration with any named cybercrime nor evaluated for suitability to practitioners' investigation needs.

In [43], the authors demonstrated the cybercriminal profiling of ID theft perpetrators. Their approach relied on a

technical analysis of the victim/offender device to uncover the intention, motivation and skills of the perpetrator. The resultant profile from the deductive method could then be used by digital forensic practitioners to speed up future investigations of similar nature. The restriction to ID theft, the technical-only analysis, and the non-incorporation into a digital forensic process methodology are the limitations to their work as far as the proposed research is concerned.

Author [40] attempted to show that profiling can be accurate and useful in the cyberspace. Social media was the experimental cyberspace from which evidence was collected for analysis using the BEA. The assertion from literature that low self-control is the best indicator of criminal propensity allowed her to focus on analysing participants' online behaviours for such traits. Two-part profiles of demographics and self-control traits were generated for all participants, and a further interview with the participants confirmed the accuracy of the profiling approach in deducing behavioural characteristics from digital evidence. However, the study was limited to deducing personality traits from online social activities rather than actual cybercriminal activities. The study's application of the BEA theory to investigate only an aspect of cyberspace behaviour also proves the necessity for a dedicated cybercriminal profiling framework for all cyber-activities.

The authors of [44] used the deductive approach to apply the behavioural evidence analysis methodology to a digital forensic framework. A cyberstalking incident was then simulated virtually to build a criminal profile, which outlined the offender characteristics, relevant evidence and their locations, as well as their connections to the offender. Their work introduced a technical aspect in addition to the usually behavioural-only analysis. However, their work was limited by the consideration of cyberstalking only, as well as the application of the behavioural evidence analysis methodology to only emails/instant messages. They also conceded that an integrated profiling methodology for digital forensic process, as well as an entirely different approach is necessary.

Recently, authors [41] applied the BEA theory to demonstrate the accuracy of cybercriminal profile for forensic investigations using 20 real cyber-incident case data from the last 5 years in Dubai. They deduced the characteristics of the offenders and victims, and also determined offender-victim relationship and the offenders' modus operandi. Their study proved that the use of BEA for profiling as a tool facilitates the identification of offender/victim's motivation and behaviour, elimination of suspects, identification of potential victims and above all, focuses direction and speed to digital forensic investigations. Although the study was limited to cyberstalking and a behavioural-only approach, it presented the need for the ad hoc application of BEA theory (a criminal profiling framework) to the digital forensic process methodology.

IV. Criminal Profiling in Traditional Investigations

Procedures employed in criminal profiling methodologies and frameworks for investigating traditional crimes typically fit within four stages, depending on the approach of the methodology. The inductive approach uses information from crime databases and victims' accounts statistically to build a criminal profile, thus, providing a generalized criminal profile from specific crime cases [28]. The deductive approach on the other hand uses detailed information about a single specific crime to build a specialized criminal profile. The inductive approach is characterized by undesirable error rates and is subject to a variety of influences including bias due to interpretation, loyalty, and time difference in cases being used, which the deductive approach is unaffected by [20]. However, the inductive approach has been mostly used than deductive approach due to the difficulty in accessing real crime case details [22]. Each profiling stage in a deductive-oriented methodology (favoured by this author) takes inputs from the crime scene and the preceding stage, and passes its outputs to the following stage [39]. The independent outputs of all stages are then combined and refined to produce a profile that would describe the unknown offender in details [45]. Traditional crimes which profiling are applicable to include murder, rape, kidnapping, theft, money laundering, stalking, trafficking, arson, arms/drugs distribution, amongst others. The following sub-sections highlight the emphatic role of profiling in solving these traditional crimes, which can be equally extended for use in cybercrime investigations.

A. Equivocal Analysis

The main aim here as the first stage is to review all sources

of evidence available in order to ensure their integrity going forward. Evidence may sometimes pose ambiguous meanings, may have been misrepresented by the first responder or lack completeness, leading to the profiler unconsciously making costly assumptions and potentially drawing inaccurate conclusions [46]. It is important to ascertain and ensure that the value of available evidence is aligned with the purpose of investigation to be conducted [20]. Physical evidence, crime scene documentation, photos, victim statements and laboratory reports are sources which usually need to be verified in order to sift irrelevant evidence away and avoid bias from the onset [47]. Other objectives of this stage include collecting all associated evidence, identifying the most relevant of the evidence for further analysis, and preparing a clear plan of procedures in the following stages [48][49].

A classical procedure in this stage is to establish the class of crime the investigation is to be based on. The mere assumption or theory by a police detective or investigator that a crime is a homicide should not be taken at face value by the profiler without being ascertained. A review of the autopsy report and photos, crime scene documentation, and perhaps a request for a sexual assault kit may reveal underlying clues to classify the supposed homicide as a rape and murder crime. Conducting the profiling based on rape in addition to murder could explain a lot more about the offender's behaviours and personalities than profiling based on murder alone.

B. Victimology

This marks the beginning of analysis of the relevant evidence available, with the main aim of inferring the offender's behaviours from their choice of victims. An analysis of the evidence is performed with the victim as the sole focus of attention in order to provide leads to understanding further details of the crime. This critical study of victim's behaviours before, during and after the attack would give a clear indication of the victim's lifestyle and situational exposure to being singled out for such attack [20].

The victim's relationship with the offender, victim's role in the attack, as well as the type of attachment, if any, are the other details that this stage uncovers [45]. These details can be used to draw inferences about the offender's motive, fantasy, expertise and modus operandi [49]. Victim statements, laboratory reports, crime scene documentation and physical evidence are sources that can be analyzed during victimology [23].

A profiling procedure instance for determining the lifestyle exposure of a murder victim would examine the laboratory reports of the victim. If a toxicology test shows a high dosage of a drug, then it could be inferred that the victim was a drug user and another user of the same strain of drug could be the offender. Another instance of a procedure in the victimology stage is to infer a situational exposure of a victim whose laboratory reports show their bowel contained a recent meal that is only served by a particular restaurant. This might mean that the victim had been exposed to their killer at the restaurant, in which case any probable drug user that was at the restaurant the same time the victim was there could be the unknown offender. It is thus evident that apart from insights to the behaviours of the probable offender, the victimology stage also helps immensely to direct focus during investigation.

C. Crime Scene Characteristics

At this stage, conscious effort is made to combine an understanding of how the offender orchestrated and committed the crime with the profiling results from the victimology stage in order to further establish the personality of the unknown offender. Analyzing specific features of a crime scene can help establish the timing, modus operandi, signature, expertise, skills/sophistication and the level of organization of the unknown offender [20][48]. The offender's choice of materials and attempt to obfuscate detection or identification The amount of clues inferable is usually dependent on the kind and extensiveness of the crime being investigated and several other details from the crime scene can describe whom the unknown offender is, what he does and where he can be found [20][28][45][48]. Although, other evidence sources may also be valuable, the physical scene/evidence, crime scene documentation and photos are the major sources for analysis in this stage [47].

An instance of a procedure in the crime scene characteristics stage is to infer the modus operandi and organization level of an offender whose victim was killed in their own house by examining the house security system for tampering, in which case non-tampering would mean that the offender convinced the victim to open the door or the offender waited and attacked at a time the door was left unlocked. A further look at the last time logged for card or key input access in the house security system would provide an idea of the time the offender perpetrated the crime.

D. Offender Characteristics

The main aim at the final stage is to infer more detailed characteristics of the unknown offender. The profiler builds on the results from victimology and crime scene characteristics to provide more details about the personality of the unknown offender [20]. The motivation and the intention of the offender are established, as well as the level of the risk the offender was ready to take to achieve their criminal objective. The mental and social disorders the unknown offender is probably afflicted with are also established at this stage [28][49]. Other details such as information about the offender's kind of work, vehicle, general demeanour, hobbies and phobias amongst others are inferred at this stage, while characteristics already established in the previous stages are then reviewed one last time before the final criminal profile is drawn [20].

The final criminal profile deduced after completing the procedures in the above stages, which could then be used to focus the direction of investigation, reduce the pool of suspects and reduce the turnaround time of the investigation, contains the following descriptive details about the presumptive offender: demographics, legal/criminal history, vocational/professional background, family background, habits and interests, personality, transport mode and residence/work location [20][45].

V. Issues in the Cyber Scene

The term cyber scene is described either as an environment in which an electronic incident has occurred or a digital environment in which a crime has been perpetrated [50][51]. Foremost, is the vast difference in the nature of the cyber scene details compared to the traditional scene. The traditional scene is a rather stable and physical one, such that reactions to direct and indirect actions with/(in) it leave real and relatively durable traces. The cyber scene in contrast has a very intangible nature, with traces from actions within it exhibiting an abstract conception and a high degree of volatility. Consequently, the elements to be examined while investigating crimes in the cyber scenes could be distorted or lost before they could be collected for use [6]. An analogy would be an attack that leaves its victim dead. The body of the victim serves as a valuable evidentiary element in a traditional scene, and is found without any special effort in most cases. However, the profile – which is the only remain of the victim from a cyber scene – usually contains limited or no evidential value, and may only be found with a deliberate and timely technical search.

Another unconventional element of the cyberspace that makes the profiling a difficult application during investigations is the variance in the number of crime scenes that exist – with no particular crime scene (pseudo-scene) or more than a single crime scene [6]. Some cybercrimes are deliberately perpetrated in environments that are even more abstract than the cyberspace, in order to make such incidents undetectable or difficult to trace back to the offender [51]. Software/movie piracy engagement, by illegally downloading, using/watching, and removing it afterwards within a virtual machine or a cloud service, would present no crime scene for the investigators. In another vein, some cybercrimes constitute more than a single crime scene when they are committed. An example is a cyberstalking crime, in which both the offender and victim's computers contain evidence of the crime, and are as such declared the crime scenes [44]. This inconsistency in the number and nature of crime scene makes the proper application of criminal profiling in such investigation difficult.

The transnationalism of cybercrimes introduces another complexity that is not found in traditional crime scenes. A typical cybercrime incident can involve an offender and victim resident in two different states, countries or continents [19]. This becomes more complex with organized

cybercrimes that could have several number of offender computer terminals attacking several victim computers using several or no proxy terminals [6]. Thus, the issues posed are the increased number and decentralization of the crime scenes that makes performing and consolidating multiple instances of the same profiling procedures on each of the crime scenes difficult, if not impossible. This is due to the fact that criminal profiling is best done when looking at all the crime scenes at the same time for the complete picture of the incident [20][50].

The legal implication of cybercrime's transnationalism is the inconsistency it brings to such investigations. Different states and countries dictate laws that consider similar or same situations differently [4][6]. This might result in a situation where full access to an out-of-jurisdiction cyber scene that partly hosted the perpetration of an organized cybercrime is denied. This would subsequently make a complete and meaningful criminal profiling impossible for the success of the investigation, since all constituents components of the crime are not being considered at the same time, as argued by [50].

Accurate cultural translation of behaviours in a traditional crime can often be the key to identifying a probable offender, especially when the investigator is a local and certain materials or methods used in committing the crime are common to that geographical area or population [20]. The use of a particular model of knife in a murder may suggest the offender is from a particular tribe within a country, who are the only population that use it. Whereas the use of the same model of knife might mean nothing, if it is a common item within the major tribes in another country. However, since cybercrimes cross geographical and cultural boundaries, investigators may encounter difficulty in accurately translating and distinguishing the appropriate interpretations within the various boundaries involved for profiling purpose [45][46].

The immaturity of the digital forensic discipline is a perpetually highlighted issue. It is generally described to be in its infancy as compared to the traditional forensic science discipline [2][52]. A few sub-challenges that have stemmed from this immaturity include the unstandardized digital forensic training/education and inadequate standardized methodologies, amongst others [4][46]. These inadequacies restrict the acquisition of the required knowledge and skills to develop and apply criminal profiling in cybercrime investigations most appropriately. It is believed that the discipline will reach its heights and demonstrate complete reliability like the traditional forensic science. However, until then, coming up with standardized methodologies and training for effective cybercriminal profiling remains a gap that needs timely address.

VI. Towards a Probable Solution

Previous researches have seen attempts to use the traditional behavioural evidence analysis methodology to profile cybercriminals. While technical profiling is as important as behavioural profiling, it has received less than equal influence on profiles built for cybercrimes. The engagement and influence of the behavioural evidence analysis methodology on the profiling procedures have been almost dominant in authors [44] and [53]'s respective deductive-oriented and inductive-oriented works. Author [32]'s unevaluated deductive-oriented methodology however, suggested very little as far as technical profiling is concerned. More importantly, each of the previous works have attempted to profile perpetrators of a single kind of cybercrime – usually cyberstalking or cyberbullying, and identity theft in [43]'s work. While there is no expectation of a single methodology to fix all the issues highlighted, it is expected that a single methodology should able to address the major issues properly. Figures 1 and 2 show the current limited approaches to transfer criminal profiling on to the cyber scene during investigations, as highlighted above.

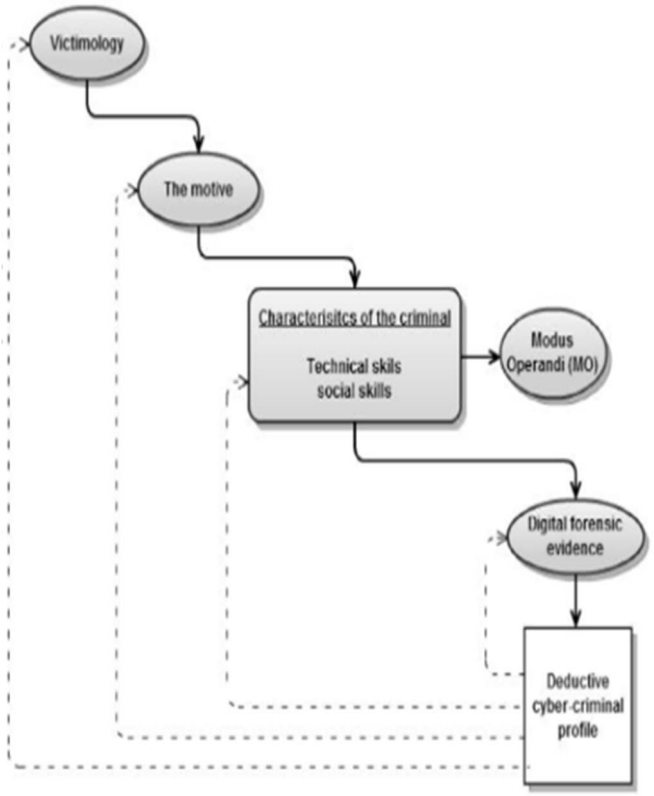


Fig. 1. Comprehensive Methodology for Profiling Cyberstalking Cybercriminals [32]

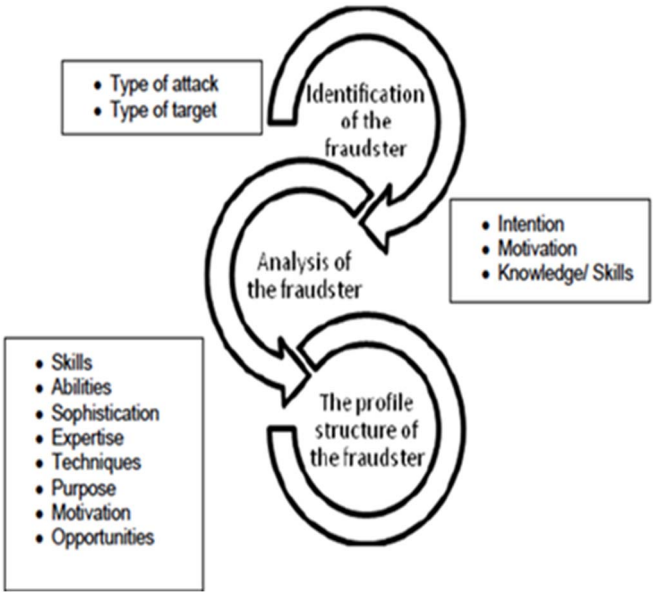


Fig. 2. Methodology for Profiling ID Theft Cybercriminals [43]

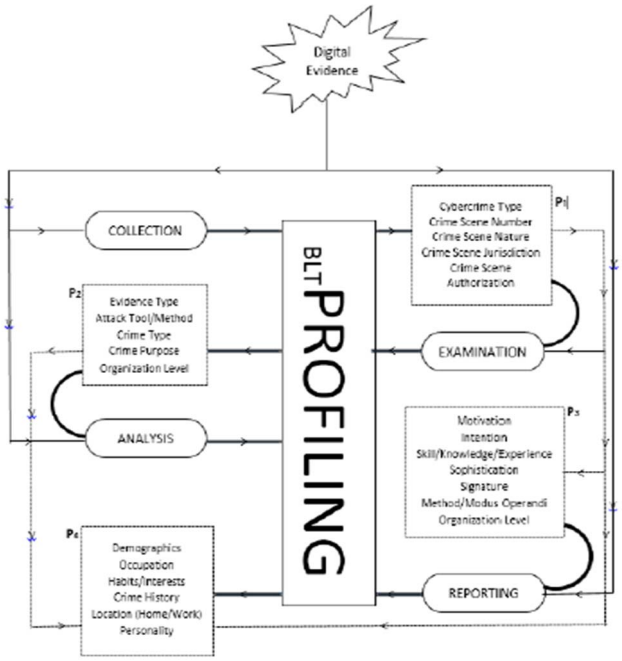


Fig. 3. Preliminary Proposed Integrated Generic Cybercriminal Profiling Framework

A. Proposed Framework

A generic, deductive-oriented profiling framework with a dominant behavioural evidence analysis influence, shown in figure 3 is thus, preliminarily recommended and proposed for future work. This framework would take an iterative approach to refining the outputs being fed into the stages from their preceding stage. It would also address multiple kinds of cybercrimes and allow the incorporation of multiple crime scenes to promote easier incident reconstruction during investigation. The proposed framework is also integrated to a digital forensics process methodology, and thus, fosters a seamless application in digital forensic investigations. In addition to the technical and behavioural facets that previous methodologies provided, a legal facet is introduced into the proposed framework to guide the process of profiling where issues of transnationalism are encountered.

VII. CONCLUSION

This paper has presented an insight into the menace that cybercrime constitutes and the obvious inadequacy in counteracting it. The role of criminal profiling as a valuable investigative technique in traditional crimes, as well as its potential contributions to enhancing cybercriminal investigations and convictions, has been highlighted. The issues responsible for the difficulty in the transference of the technique to cybercrime investigations have been discussed, and a preliminary framework that builds on previous solutions has been proposed. The details of the proposed solution would be discussed in subsequent publications by the authors.

REFERENCES

[1] Parliamentary. (2006). Computer crime. *Postnote* [Online]. Available: http://www.parliament.uk/documents/post/postpn271.pdf

[2] M. Sheetz, Computer Forensics: an essential guide for accountants, lawyers, and managers. Hoboken: John Wiley & Sons, 2007.

[3] J. Williams, ACPO Good Practice Guide for Digital Evidence, 2012.

[4] S. Garfinkel, "Digital forensics research: The next 10 years," *Digital investigation*. 7. p.64-73. 2010.

[5] M. Solomon, K. Rudolph, E. Tittel, N. Broom, and D. Barrett, Computer Forensics Jumpstart. 2nd ed.: Wiley / Sybex, 2011.

[6] E. Casey, Digital evidence and computer crime: forensic science, computers and the internet. Oxford: Academic Press, 2011.

[7] G. Mohay, A. Anderson, B. Collie, O. D. Vel, and R. Mckemmish, Computer and Intrusion Forensics. Artech House inc., 2003.

[8] FBI, F. B. O. I., "2015 Internet Crime Report," [Online], Available: https://pdf.ic3.gov/2015_ic3report.pdf, 2016.

[9] PI, P. I., "Cost of data breach study: Global analysis," [Online], Available: http://www-03.ibm.com/security/data-breach, 2016.

[10] J. Moar, "The future of cybercrime & security: Financial & corporate threats & mitigation 2015-2020," 2015.

[11] I. Kessem, "2016 cybercrime reloaded: our predictions for the year ahead," Available: securityintelligence.com/2016-cybercrime-reloaded-our-predictions-for-the-year-ahead, 2016.

[12] Statista, "Cyber crime: Largest online data breaches," [Online]. Available: https://www.statista.com/statistics/290525/cyber-crime-biggest-online-data-breaches-worldwide, 2016.

[13] J. Cox, "Another day, another hack: User accounts of dating site Badoo" [Online]. Available: motherboard.vice.com/read/another-day-another-hack-user-accounts-of-dating-site-badoo, 2016.

[14] BBC, B. B. C. "Attack on yahoo hit 600 million users," [Online] Available: bbc.com/news/world-us-canada-37447016?ns_mchannel=social&ns_campaign=bbc_breaking&ns_source=twitter&ns_linkname=news_central, 2016.

[15] J. Leyden, "Megabreach: 55 million voters' details leaked in philippines," Available: theregister.co.uk/2016/04/07/philippine_voter_data_breach, 2016.

[16] Mybroadband, "Ethekwini municipality leaked private details of almost 100,000 residents," [Online] Available: mybroadband.co.za/news/security/179064-ethekwini-municipality-leaking-private-details-of-over-300000-residents.html, 2016.

[17] E. Casey, *Handbook of digital forensics and investigation*. Oxford: Academic Press, 2009.

[18] N. R. Council, D. E. P. Sciences, C. A. T. Statistics, P. G. Affairs, T. L. Committee on Science, & C. I. N. F. S. Community, "*Strengthening forensic science in the united states: a path forward,*" National Academies Press, 2009.

[19] A. Marshall, Digital Forensics: Digital Evidence in Criminal Investigations, 2008.

[20] B. Turvey, Criminal Profiling: An Introduction to Behavioural Evidence Analysis. 4th ed. Sitka: Elsevier Ltd., 2011.

[21] A. Lowe, "Criminal profiling in the investigative process," National Legislative Eagle, vol. 8, pp. 6, 2002.

[22] D. Canter, "Offender profiling and investigative psychology," *Journal of investigative psychology and offender profiling. vol.* 1, pp. 1-15, 2004.

[23] B. Turvey, Criminal Profiling: An Introduction to Behavioural Evidence Analysis. Oxford: Academic Press, 2008.

[24] HO, H. O., "Cybercrime: A review of the evidence, summary of key findings and implications," Available: https://gov.uk/government/uploads/system/uploads/attachment_data/file/246749/horr75-summary.pdf, 2013.

[25] USDOJ, D. O. J., "The Federal Bereau of Investigation's ability to address the national security cyber intrusion threat," Available:http://oig.justice.gov/reports/fbi/a1122r.pdf, 2011.

[26] NCRB, N. C. R. B. "Crime in India: 2015 Statistics," Available: http://nrcb.gov.in, 2016.

[27] J. Gaur, "Crime conviction imbalance and the use of forensic science evidence," *Indian police journal*, 2005.

[28] M. Rogers, "The Role of criminal Profiling in the Computer Forensics Process," Computer Security. vol. 22, pp. 292-298. 2003.

[29] S. Ingram, "If the Profile Fits: Admitting Criminal Psychological Profiles into Evidence in Criminal Trials," Journal of Urban and Contemporary Law. vol. 54. pp. 239-266, 1998.

[30] W. Bennett and K. Hess, "Investigating Violent Crimes in Criminal Investigation," Cengage Learning, 2007, pp. 296-313.

[31] N. Parsons-Pollard and L. Moriarty, "Cyberstalking: Utilizing What We do Know, Victims & Offenders," An International Journal of Evidence-based Research, Policy and Practice. pp. 435-441, 2009.

[32] H. Tennakoon, "The need for a comprehensive methodology for profiling cyber-criminals," Available: http://www.newsecuritylearning.com/index.php/feature/150-the-need-for-a-comprehensive-methodology-for-profiling-cyber-criminals, 2012.

[33] J. Barnes, "Technological Profiling of Cyber-Bullying," [Online] Available: http://www.dfinews.com/articles/2013/09/technological-profiling-cyber-bullying, 2013.
[34] K. Mandia, C. Prosise, and M. Pepe, Incident response and computer forensics (2nd ed.) McGraw Hill, California, 2003.
[35] J. Douglas, R. Ressler, A. Burgess, and C. Hartman, "Criminal profiling from crime scene analysis," *Behavioral sciences and The Law. vol. 4*, pp. 401-421, 1986.
[36] J. Jackson and D. Bekerian, Offender Profiling: Theory, Research and Practice. Wiley, 1997.
[37] I. Alison, C. Benell, A. Mokros, and D. Ormerod, "The personality paradox inoffender profiling: a theoretical review of the processes involved in deriving background characteristicsfrom crime scene actions," *Psychology, public policy, and law.* vol. 8, pp. 115-135, 2002.
[38] J. Beasley, "Serial murder in America: Case studies of seven offenders," *Behavioral Sciences & The Law. vol.* 22, pp. 395-414, 2004.
[39] D. Canter, I. Alison, E. Alison, and N. Wentink, "The organized/disorganized typology of serial murder: myth or model?," *Psychology, public policy, and law. vol.* 10, pp. 293-320, 2004.
[40] Y. Szde, "Behavioural Evidence Analysis on Facebook: A Test of Cyber-Profiling," 2013.
[41] N. A. Mutawa, J. Bryce, V. Franqueira, and A. Marrington, "Forensic investigation of cyberstalking cases using behavioural evidence analysis," *Digital investigation.vol.* 16, pp. 96-103, 2016.
[42] A. Stanojoska, B. Petrevski, and K. Nikolova, "How a cybercriminal does look like? Building a profile of modern criminals," 2012.
[43] O. Angelopoulou, I. Vidalis, and I. Robinson, "Who are you today? Profiling the id theft fraudster," 2013.
[44] A. Silde and O. Angelopoulou, "A digital forensics profiling methodology for the cyberstalker," *Proceedings of international conference on intelligent networking and collaborative systems, IEEE INCoS, p*p. 445-450, 2014.
[45] R. Kocsis, Criminal Profiling: Principles and Practice, USA: Humana Press Inc., 2006.
[46] E. Casey, "The value of behavioral analysis in digital investigations," Digital Investigation, vol. 3, pp. 57, 2006.
[47] W. Petheerick, B. Turvey, and C. Ferguson, Forensic Criminology, 1st ed. London: Academic Press, 2009.
[48] J. Douglas and L. Douglas, "Modus Operandi and the Signature Aspects of Violent Crime," *In:* Crime Classification Manual: A Standard System for Investigating and Classifying Violent Crimes, J. Douglas, A. Burgess, and R. Ressler, Eds. San Francisco: Jossey-Bass, 2006.
[49] C. Ferguson and B. Turvey, "Victimology: A brief history with an introduction to forensic victimology," *In:*Forensic victimology: Examining violent crime victims in investigative and legal contexts, B. Turvey and W. Petherick, Eds. 2009.
[50] J. Vacca, Computer Forensics. Higham: Charles River Media, 2002.
[51] S. Garfinkel, "Anti-Forensics: Techniques, Detection and Countermeasures," [Online] Available: http://calhoun.nps.edu/bitstream/handle/10945/44248/Garfinkel_Anti-Forensics_2007.ICIW.AntiForensics.pdf?sequence=1, 2007.
[52] M. Meyers and M. Rogers, "Computer Forensics: The need for standardization and certification," International Journal of Digital Evidence, vol. 3, pp. 1-11, 2004.
[53] M. Rogers, "Psychological profiling as an investigative tool for digital forensics," In Digital Forensics:Threatscapes and Best Practices, J. Sammons, Eds. New York: Syngress, 2015, pp. 45-57.

Mitigation of Impulse Noise in Powerline Systems using ANFIS Technique

Olamide M. Shekoni, Ali N. Hasan and T. Shongwe
Department of Electrical &Electronic Engineering
University of Johannesburg
South Africa
olamide.shekoni@yahoo.com, alin@uj.ac.za, tshongwe@uj.ac.za

***Abstract*—The use of OFDM channel for the transmission of data in power line communication (PLC) system has been of several importance to technology development. However, during transmission, the OFDM channel is greatly disturbed by impulse noise that causes a wrong information to be received. Several techniques such as iteration, coding, clipping and nulling methods have been used to lessen the upshot of impulse noise in OFDM channel. However, these techniques still suffer some drawbacks and require a high signal-to-noise (SNR) power for high performance. This paper presents an advanced use of artificial neuro-fuzzy inference system (ANFIS) technique in removing the complete impulse noise and some of the additive white Gaussian noise (AWGN) that were mixed with the transmitted data in an OFDM channel and using the minimum SNR power. Obtained results propose that ANFIS technique can be used to mitigate impulse noise from a powerline communication channel.**

Keywords—Machine learning, Orthogonal frequency division multiplexing, additive white Gaussian noise, impulse noise, encoders, ANFIS, bit error rate, power line communication

I. INTRODUCTION

Power-line communication (PLC) system is commonly used to transmit data and signals from the transmitter to the receiver end. PLC has also been used for several technology purposes such as home automation, smart grid metering and internet access referred to as "broadband over powerline (BPL)" [1]. Some of the merits of PLC include low cost of implementation, easy to install, serves well for distribution purposes and can work with an existing electrical wiring [2]. However, data transmission through PLC is greatly disturbed by both impulse noise and background noise [3]. Other sources of distortion in PLC include frequency-selectivity and high-channel attenuation [4]. Both impulse noise and background noise contribute majorly to the high packet rate or bit error rate (BER) in powerline systems [5]. The impulse noise varies rapidly with time and can be classified into asynchronous impulse noise (AIN), periodic impulse noise synchronous to main frequency and periodic impulse noise asynchronous to the main frequency [6, 7]. Background noise varies slowly with change in time. Background noise is classified into narrowband noise and coloured noise. A good example of coloured noise is the additive white Gaussian noise (AWGN) [8]. Several schemes of suppressing and eliminating impulse noise (IN) were studied and can be classified into two groups namely parametric and non-parametric method. The Parameter methods aim to obtain the optimal onset to process the received signal using parameter estimation with the noise model. This method is known as nonlinear methods. Parametric methods try to seek a maximum threshold that limits the power of impulse noise. Regrettably, parametric methods still interfere with the signal in the process of limiting the impulse noise. Furthermore, parametric method involves extra training overhead and once the noise model mismatches the time varying noise statistics, they suffer degradation in the performance. Non-parametric methods however can bypass the demand of previous knowledge and noise model by using the sparse arrangement of impulse noise in the time domain [9].

Recent work is done using clipping and nulling (blanking) technique, interleaves, iterative and coding techniques to reduce the effect of impulse noise from a corrupted data in PLC channel [9, 10, 11]. However, the use of clipping requires a good understanding of the corrupted signal and determining the corrupt signal magnitude in order to proffer a threshold value (T_h) and nulling (setting to zero) that tends to remove the impulse noises from the corrupted signal. Equation (1) explains the mathematical process of clipping while equation (2) presents the nulling methodology,

$$\vec{r}_k = \begin{Bmatrix} r_k, \text{ for} |r_k| \le T_h \\ T_h e^{j \arg(r_k)}, \text{ for} |r_k| \ge T_h \end{Bmatrix} \rightarrow \textit{Clipping phase} \quad (1)$$

$$\vec{r}_k = \begin{Bmatrix} r_k, \text{ for} |r_k| \le T_h \\ 0, \text{ for} |r_k| > T_h \end{Bmatrix} \rightarrow \textit{Nulling phase} \quad (2)$$

Where r_k is the transmitted signal mixed with impulse noise and AWGN noise while $\vec{r}_k$ is the received sample after clipping and nulling process has been performed to remove the impulse noise from the transmitted signal (r_k) [12, 13].

Iterative technique is an estimation technique and can be implemented by subtracting an estimated value of the impulse noise (IN) from the received signal vector (r) [14]. However, this technique demands a high number of iterations for performance improvement. This method can be time consuming and does not necessarily remove the impulse noise completely [15]. While the use of error correcting codes such as convolution coding and Reed-Solomon (RS) coding are relatively easy to implement, present a better mitigation

performance [16, 17]. However, some of these error-correcting techniques suffer some limitations [18, 19]. For example, RS code performance is poor with binary-phase-shift-keying (BPSK) modulator while convolution codes are suitable for linear-time invariant (LTI) systems and underperforms with non-linear LTI systems [20, 21].

The contribution of this paper is to introduce an advanced use of artificial neuro-fuzzy inference system (ANFIS) for the extenuation of corrupt impulse noise present in an orthogonal frequency division multiplexing (OFDM) channel.

The layout of this paper is organized as follows, section 2 will present a summary of the used ANFIS technique. In section 3, a report of the experiment setup and method is provided. Section 4 will present the results, and section 5 will include the conclusions.

II. ANFIS TECHNIQUES

This is a supervised machine learning technique that combines both Sugeno (a type of fuzzy logic control inference system) and artificial neural network (ANN) into a single method recognized as artificial neuro-fuzzy inference system (ANFIS) technique [22]. In an ANFIS controller, ANN is used to automatically regulate the membership functions (MFs) using backpropagation alone or combined with least square technique to lower the level of inaccuracies in the determination of rules in fuzzy logic systems. [23] Merits of ANFIS include easy implementation for both linguistic and numeric knowledge, fast and accurate learning, and does not need prior human knowledge. Applications of ANFIS include optimization of energy from a photovoltaic system, prediction of lung detection risk in humans, edge detection and Arabic alphabet detection [24]. However, a major demerit of ANFIS is the limitation to one output variable [25, 26, 27].

III. SIMULATION MODEL

To examine the practicability of the proposed noise cancellation algorithm with the use of ANFIS technique to adaptively distinguish and moderate the impulse noise present in the transmitted signal (T_X) in an OFDM system, an experiment was conducted by modelling a first-order Sugeno FIS (fuzzy inference system) using some set of rules (e.g. IF, AND, and NOT). A complete OFDM system that comprises of bernoulli signal generator for generating the transmitted signal (T_x), binary phase shift keying (BPSK) modulator and demodulator for converting the bit samples (0 and 1) into complex values and vice versa, Inverse-fast-fourier-transform (IFFT), FFT (fast fourier transform), BPSK demodulator, AWGN channnel and with random impulse noise generator were implemented to achieve the objectives of the experiment. The random impulse noise data were generated using equations (3-5),

$$F_m(n_k) = \sum_{m=0}^{\infty} p_m \mathbb{N}(n_k; 0, \sigma_m^2) \tag{3}$$

$$\text{where,} \quad P_m = \frac{A^m e^{-A}}{m!} \tag{4}$$

$$\text{and} \quad \sigma_m^2 = \sigma_i^2 \frac{m}{A} + \sigma_g^2 = \sigma_g^2 \left(\frac{m}{A\Gamma} + 1 \right) \tag{5}$$

$F_m(n_k)$ is the probability density distribution (PDF) of a noisy signal (n_k), $\mathbb{N}(n_k; \mu, \sigma_m^2)$ denotes the Gaussian PDF, μ is the mean, σ^2 is the variance with k -samples, σ_i^2 is the impulse noise variance, σ_g^2 is the AWGN (background noise) and Γ denotes the ratio between Gaussian power (σ_g^2) and impulse noise power (σ_i^2) and is computed mathematically as $\frac{\sigma_g^2}{\sigma_i^2}$. The parameter (A) is the impulse density.

Figure 1 displays the magnitude of the impulse noise. It illustrates the amplitude value of the corrupt impulse noise that got mixed with the transmitted data (T_x) added for a duration of 500 seconds.

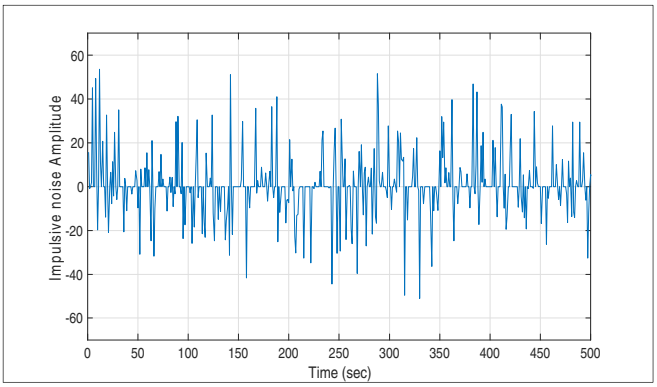


Fig. 1. Impulse signal Magnitude

Figure 2 presents the block illustration of a complete OFDM system developed using two ANFIS controllers. In the trasmitter, the generated bernoulli signal is modulated using BPSK modulator and channeled through IFFT. Then the output data from the IFFT passess through AWGN channel, through which AWGN is been added to the transmitted signal and subsequently, impulisve noise (v) is also added to the transmitted signal (t). This addition goes on to make up the corrupted transmitted signal (u) as seen in the figure. The ANFIS controllers were trained using the 10,000 input samples of signals u and v as the input variables (predictors) and signal t (output signal from the IFFT/AWGN block) as the target.

For simplicity, the complex-variable data $u = u_1 + ju_2, v = v_1 + jv_2,$ and $\vec{t} = \vec{t_1} + j\vec{t_2}$ were split into real and imaginary components and both the real and imaginary parts of signals (u, v) were the ANFIS controllers inputs (u_1, u_2, v_1, v_2) while the real and imaginary component of signal t (t_1, t_2) were used as targets to predict responses $(t_1{*}, t_2^{*})$ as the output from the ANFIS 1 and ANFIS 2 controller respectively. This output is then demodulated, passed through the FFT channel and compared with the transmitted signal.

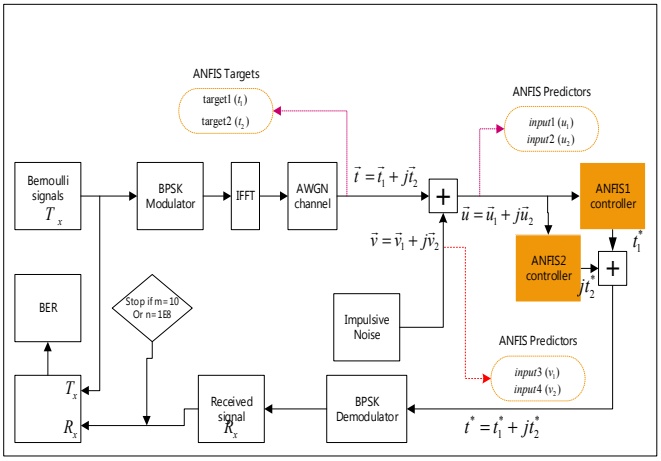


Fig. 2. Complete OFDM channel Using ANFIS Techniques

Figure 3 shows the flowchart algorithm of the used ANFIS error-correcting technique.

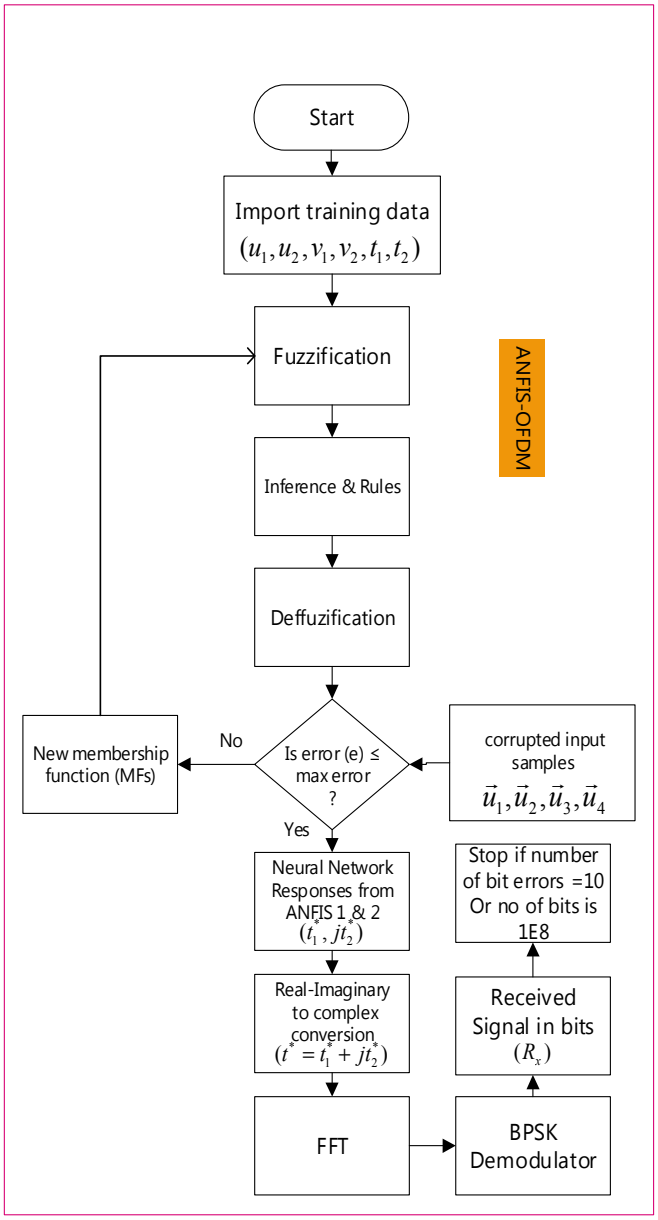


Fig. 3. ANFIS Algorithm

IV. EXPERIMENTAL RESULTS

Figure 4 gives a comparison of the ANFIS-treated OFDM, theoretical OFDM (OFDM corrupted by AWGN only), and highly corrupt OFDM channel (OFDM corrupted with AWGN and impulse noise as the uncorrected OFDM within the range of 0 dB to 40 dB noise power. The outcome of the experiment displays the high performance of the ANFIS impulse-noise error correcting technique as the bit-error-rate (BER) with ANFIS was the lowest in most of the considered signal-to-noise (SNR) levels (0 dB, 4 dB, 8 dB, 10 dB, and 12 dB) and using the lowest SNR power (dB) where the variace of the noise was set to 50 and probability of impulse noise present was 50%.

After removing the entire impulse noises that were mixed with the transmitted data, it can also be observed that the ANFIS technique displays the capability of removing some of the AWGN channel noise from the OFDM channel. Similarly, the variance of the impulse noise was varied at a fixed value of probabilty of impulse noise present. The value of the variance of the impulse was also fixed while the probability was varied and the improvement of the PLC was observed, the BER perfomances for these are displayed in Figures (5-12).

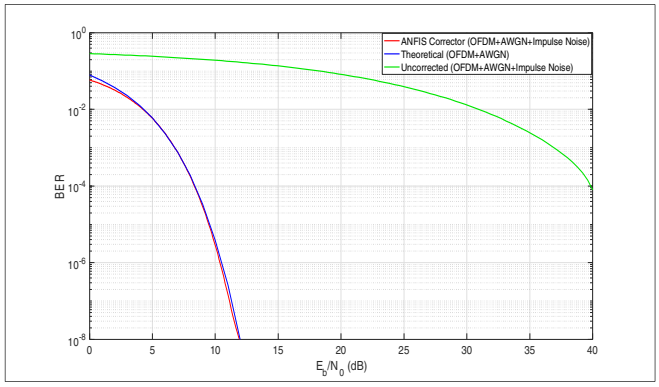


Fig. 4. Comparion of the bit rror rate of theoretical OFDM with uncorrceted IN-corrupted OFDM and ANFIS-treated OFDM, with impulse noise of varance (V) = 0.5 and probability (P) = 0.5

Fixing the variance of the impulse noise at 50, and varying the probabilty of the impulse noise, the following graphs were obtained.

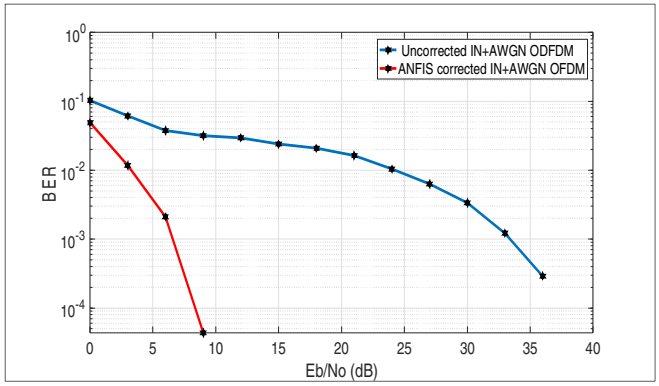


Fig. 5. BER graph with variance IN(V) = 50 probability (P) = 10%

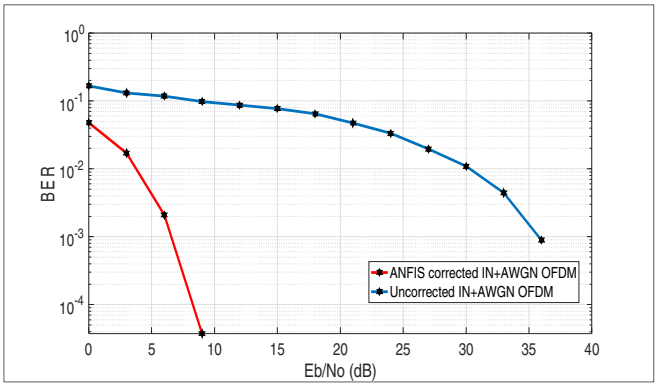


Fig. 6. BER graph with variance of IN (V) = 50 probability (P) = 30%

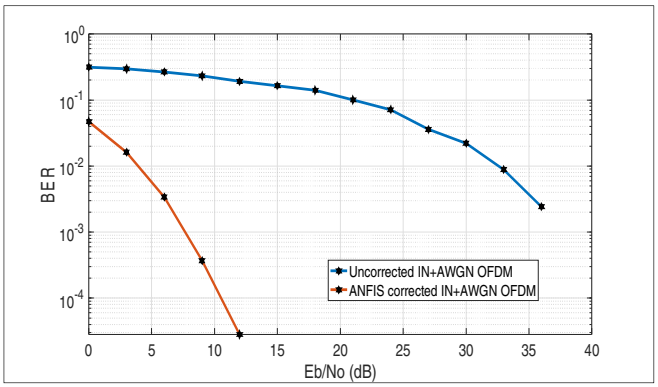


Fig. 7. BER graph with variance of IN (V) = 50 probability (P) = 70%

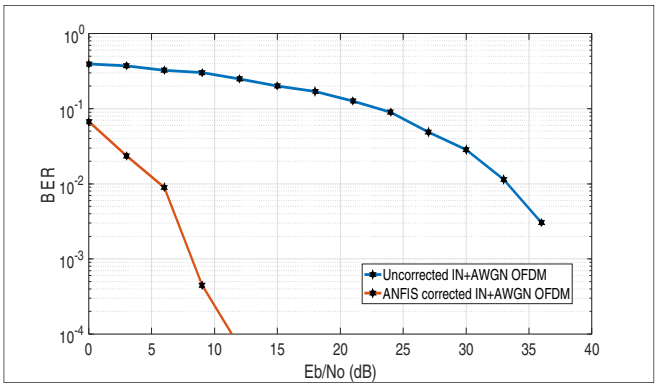


Fig. 8. BER graph with variance of IN (V) = 50 probability (P) = 90%

Fixing the probability of the impulse noise at 50%, and varying the variance of the impulse noise, the following graphs were obtained.

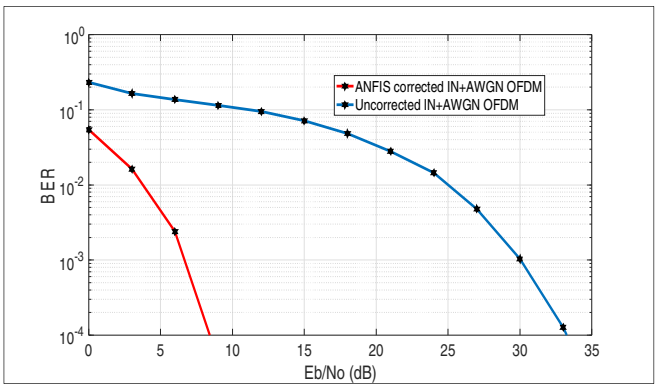


Fig. 9. BER graph with probability of IN (P) = 50% variance (V) = 10

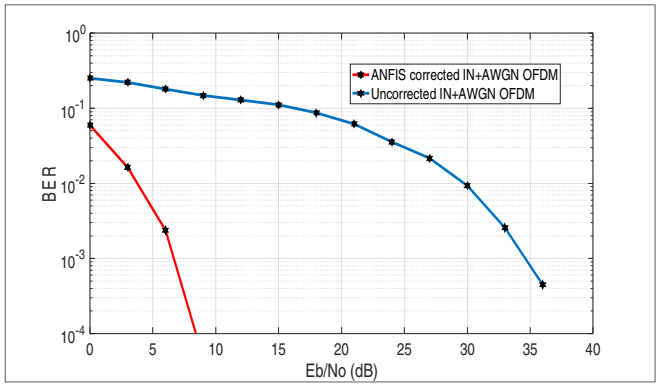


Fig. 10. BER graph with probability of IN (P) = 50% variance (V) = 30

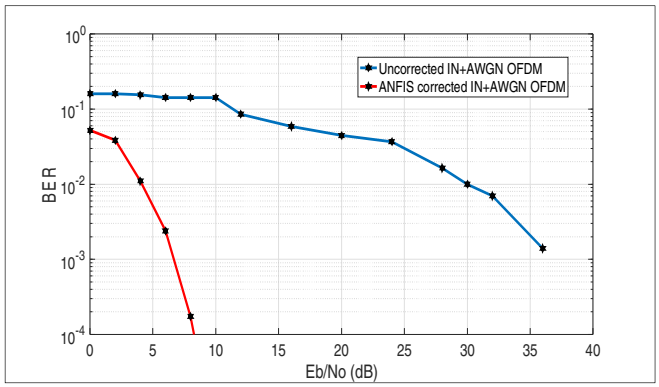


Fig. 11. BER graph with probability of IN (P) = 50% variance (V) = 50

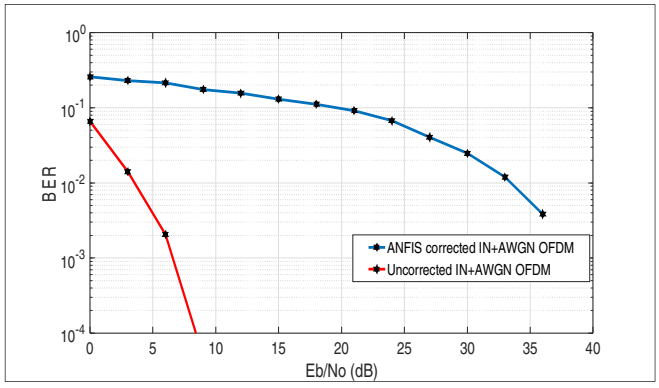


Fig. 12. BER graph with probability of IN (P) = 50% variance (V) = 70

Figures 5 to 12 illustrates the BER performance of the signal for uncorerected corrupt signal and the ANFIS corrected signal. For figures 5 to 8, the variance of the impulse noise was fixed at 50, while the probability of the noise was varied to the values of 10%, 30%, 70% and 90%.

It can be observed from Figures 5 to 8 that at lower probability values, the BER of the ANFIS corrected signal improved better (<10dB) unlike at higher probability (70% & 90% impulse noise) where the SNR value was slightly higher than 10dB. Also considering figures 9 to 12, where the probability of the impulse noise is fixed at 0.5, and the variance changed between 10, 30, 50 and 70, it is seen that these changes in the variance has little or no effect in the ANFIS result. In general, the efficiency of the implementation of ANFIS in eliminating impulse noise in PLC signal was confirmed.

V. CONCLUSIONS

This article presents an advanced use of ANFIS technique in removing the complete impulse noise and some of the AWGN that were mixed with the transmitted data in an OFDM channel using minimal signal-to-noise ratio (SNR) power. However, it can be observed from the graphs that the ANFIS performance is not affected by change in variance of the impulse noise (IN). However, it is slightly affected by the change in probability (P) of IN. As seen in Figures 6 and 7, when the probability of IN increased, the performane of ANFIS slightly degraded to SNR of about 12 dB at BER of 10^{-5} , compared to the SNR of about 9dB for P = 0.3. Obtained results advise that ANFIS technique can be recommended in mitigating impulse noise in a powerline communication channel.

REFERENCES

[1] C. Cano, A. Pittolo, D. Malone, L. Lampe, A. M. Tonello, A. G. Dabak, "State of the art in power line communications: From the applications to the medium," IEEE Journal on Selected Areas in Communications, vol. 34, no. 7, pp. 1935-1952, 2016.

[2] D. B. Ünsal and T. Yalçınöz, "Applications of new power line communication model for smart grids," International Journal of Computer and Electrical Engineering, vol. 7, no. 3, pp. 168-178, 2015.

[3] L. Bai, M. Tucci, S. Barmada, M. Raugi, and T. Zheng, "Impulsive Noise Characterization in Narrowband Power Line Communication," 11, pp. 863-880, 4.

[4] N. M. Manikanda and M. Chandrasekaran, "An efficient and low power impulsive noise suppressor architecture for OFDM system," Cluster Computing (Springerlink), vol. 2018, pp. 1-7, 2018.

[5] J. Radić and N. Rožić, "Adaptive impulse noise suppression in OFDM systems," Elektronika, vol. 1, no. 94, pp. 1-6, 2010.

[6] T. Shongwe, A. J. Han Vinck, and H. C. Ferreira, "On Impulse Noise and its Models," in 18th IEEE International Symposium on Power Line Communication and its Applications (ISPLC), Glasgow, UK, 2014.

[7] X. Zhang, Z. Yuping, and L. Zou, "An efficient impulsive noise mitigation scheme for over-sampled OFDM systems," IEEE Transactions on Consumer Electronics, vol. 55, no. 2, pp. 360-365, 2009.

[8] S. V. Zhidkov, "Impulsive noise suppression in OFDM-based communication systems," IEEE Transactions on Consumer Electronics, vol. 49, no. 4, pp. 944-948, 2003.

[9] M. R. Ahadiat, P. Azmi, and A. Haghbin, "Impulsive noise estimation and suppression in OFDM systems over in-home power line channels," International Journal of Communication Systems, vol. 2014, no. 30, pp. 1-15, 2014.

[10] K. OGURA, A. Nakamura, and M. Itami, "A study on class A impulsive noise cancellation and channel estimation under Rayleigh fading Environment," in 2018 IEEE International Conference on Consumer Electronics (ICCE), Hue City, Vietnam, 2018.

[11] S. Deepa and S. Kuzhaloli, "A survey on interleavers used for OFDM systems on various applications," International Journal of Pure and Applied Mathematics (IJPAM), vol. 118, no. 17, pp. 499-507, 2018.

[12] T. Shongwe, A. J. Han Vinck, and H. C. Ferreira, "The effects of periodic impulsive noise on OFDM," in 2015 International Symposium on Power Line Communications and its Applications (ISPLC), Austin, TX, USA, 2015.

[13] R. Chol Ho, S. Eisenbarth, and Y. Chon Ri, "Performance evaluation of OFDM transmission over impulsive noisy channels," in 2017 IEEE International Conference on Computer, Information and Telecommunication Systems, Dalian, China, 2017.

[14] J. Zhang, H. Zhiqiang, P. Chen, and Y. Rong, "A compressive sensing Based iterative algorithm for channel and impulsive noise estimation in underwater acoustic OFDM systems," in OCEANS 2017 - Anchorage, Anchorage, AK, USA, 2017.

[15] A. Al-Dweik, A. Hazmi, B. Sharif, and T. Charalampos, "Efficient Interleaving Technique for OFDM System over Impulsive Noise Channels," in 28th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Istanbul, Turkey, 2010.

[16] Y. Himeur and A. Boukabou, "OFDM-based power-line communication enhancement using a turbo coded adaptive impulsive noise compensator," Telecommunication Systems (Springerlink), vol. 2015, pp. 1-14, 2015.

[17] C. Anusuya, M. Jothimani, and R.Kalaivani, "OFDM based power line communication intensification using RSC coding with adaptive noise compensator," in International Conference on Emerging trends in Engineering, Science and Sustainable Technology (ICETSST-2017), Erode, India, 2017.

[18] "Practical implementation of adaptive analog nonlinear filtering for impulsive noise mitigation," in 2018 IEEE International Conference on Communications (ICC), Kansas City, Kansas, USA, 2018.

[19] A. N. Hasan and T. Shongwe, "Proposed machine learning system to predict and estimate impulse noise in OFDM communication system," in 42nd Annual Conference of the IEEE Industrial Electronics Society, IECON 2016, Florence, Italy, 2016.

[20] P. Chen and Y. Rong, "Joint channel estimation and impulsive noise mitigation in underwater coustic OFDM communication systems," IEEE Transactions on Wireless Communications, vol. 16, no. 9, pp. 6167-6178, 2017.

[21] A. N. Hasan and T. Shongwe, "Impulse noise detection in OFDM communication system using machine learning ensemble algorithm," in Advances in Intelligent Systems and Computing, San Sebastian, Spain, Springerlink, 2017, pp. 1-10.

[22] A. M. Farayola, Comparative study of different photovoltaic MPPT techniques under various weather conditions (Thesis submitted in partial fulfillment for the degree of Masters in Electrical Engineering), Johannesburg, South Africa: University of Johannesburg, 2017.

[23] A. M. Farayola, A. N. Hasan, A. Ali, and B. Twala, "Distributive MPPT approach using ANFIS and perturb&observe techniques under uniform and partial shading conditions," in International Conference on Artificial Intelligence and Evolutionary Computations in Engineering Systems & Power, Circuit and Information Technologies (ICPCIT-2017), India, Springer, Singapore, 2018, pp. 27-37.

[24] A. M. Farayola, A. N. Hasan, and A. Ali, "Curve fitting polynomial technique compared to ANFIS technique for Maximum Power Point Tracking," in 8th IEEE International Renewable Energy Congress (IREC), Amman, Jordan, 2017.

[25] A. M.Farayola, A. N. Hasan, and A. Ali, "Optimization of PV systems using data mining and regression learner MPPT techniques," Indonesian Journal of Electrical Engineering and Computer Science (IJEECS), vol. 10, no. 3, pp. 1-9, 2018.

[26] M. Zimmermann and K. Dosert, "Analysis and modelling of impulsive noise in broadband powerline communication and visible light communication," IEEE Transaction on Electromagnetic, vol. 44, pp. 249-258, 2002.

[27] Y. Himeur and A. Boukabou, "An efficient impulsive noise cancellation scheme for power-line communication systems using ANFIS and chaotic interleaver," Digital Signal Processing, vol. 66, pp. 1-14, 2017.

Detecting and Removing the Impulsive Noise in OFDM Channels Using Different ANN Techniques

Olamide M. Shekoni, Ali N. Hasan and T. Shongwe
Department of Electrical &Electronic Engineering,
University of Johannesburg, South Africa
olamide.shekoni@yahoo.com, alin@uj.ac.za, tshongwe@uj.ac.za

Abstract---**Orthogonal frequency division multiplexer (OFDM) is a recent modulation scheme used to transmit signals across power line communication (PLC) channel due to its robustness against some known PLC problems. However, this scheme is greatly affected by the impulsive noise (IN) and often causes corruption with the transmitted bits. Different impulsive noise error correcting methods have been introduced and used to remove impulsive noise in OFDM systems. However, these techniques suffer some limitations and require much signal to noise ratio (SNR) power to operate. In this paper, an approach of designing an effective impulsive-noise error-correcting technique was introduced using three-known artificial neural network techniques (Levenberg-Marquardt, Scaled conjugate gradient, and Bayesian regularization). Findings suggest that both Bayesian regularization and Levenberg-Marquardt ANN techniques can be used to effectively remove the impulsive noise present in an OFDM channel and using the least SNR power.**

***Keywords—Artificial Neural Network, Bayesian Regularization, Bit error rate, Binary Phase Shift Keying, Levenberg-Marquardt, Machine learning, Scaled-conjugate, Clipping and Nulling, Power Line Communication*.**

1. INTRODUCTION

Orthogonal frequency division multiplexer (OFDM) is a developed modulation technique that is recently used in powerline communication (PLC) due to its robustness against some known PLC challenges such as frequency-selective fading, multipath and interference [1]. An OFDM system is a multicarrier communication technique that uses both inverse fast Fourier transform (IFFT) and fast Fourier transform (FFT) for modulation and demodulation respectively. Industrial application of OFDM includes field-programmable gate array (FGPA) that is commonly used for high performance computing, communication and broadcast [2].

Despite its merits, OFDM is still disturbed by some noise signals such as impulsive noise (IN) that spreads among its subcarriers during transmission. This impulsive noise imposes the risk of data corruption at the receiver end and increases when the energy present in the impulsive noise exceeds a certain threshold (background noise level) [3, 4]. Hence, the need for an impulsive-noise removal technique is essential in an OFDM system. The impulsive noise is categorized into three: periodic impulsive noise synchronous to the main frequency, asynchronous impulsive noise (AIN), and periodic impulsive noise asynchronous to the main frequency [5, 6]. The periodic impulsive noise synchronous with the mains is a cyclostationary noise that commonly exists in silicon-controlled rectifiers (SCR) power supply and operates at certain frequency (e.g. 50 Hz or 100 Hz in European countries) [7, 8]. AIN is a form of noise that occurs rapidly due to ON and OFF of electrical devices while periodic impulsive noise asynchronous to the main frequency is a type of noise similar to the synchronous impulsive noise but operates at a frequency [8].

Different error-correcting techniques have been proposed and used to remove impulsive noises from OFDM channel. However, most of these techniques suffer some limitations. For example, clipping and nulling (blanking) technique was used to remove impulsive noise [9]. However, this technique demands a good knowledge of the impulsive noise magnitudes and predicting a clipping threshold that only removes impulsive noise below the clipping threshold [10]. Similarly, iterative technique, an impulsive-noise correcting method that works with the difference between the impulsive noise (IN) and the received signal vector (r). Iterative technique requires a high number of iterations for performance improvement, time consuming, and does not eliminate the complete impulsive noises from the PLC channel [11]. Lastly, the use error correcting codes such as turbo, convolution coding and Reed-Solomon (RS) code and low parity check coding that exhibits a high performance in removing impulsive noise in PLC suffers different shortcomings [12]. For instance, convolution code performs poorly with BPSK modulator (a component in OFDM channel) while convolution codes are unsuitable for non-linear time invariant (NLTI) systems [13].

The contribution of this paper is to introduce an innovative use of artificial neural network (ANN) for the mitigation of impulsive noise from an OFDM channel. Second contribution is a work done using three different ANN optimization techniques (Levenberg-Marquardt, scaled conjugate, and Bayesian ANN) to determine a more fitting ANN method that can be considered for impulsive noise mitigation in OFDM systems.

The structure of this paper is arranged as follows, section 2 will present a summary of the used ANN techniques. In section 3, a report of the experiment setup and method is provided. Section 4 will present the results, and section 5 will include the conclusions.

II. ANN TECHNIQUES

Artificial neural network (ANN) technique is a machine learning technique that is commonly used to solve some worldly (both linear and non-linear) problems due to its fast computation time and accuracy. ANN incorporates the use of activation functions (e.g. tansig function) for computation task and the interconnection of several hidden layers and neurons to link the inputs and outputs (targets) of the network together [14]. ANN can be classified into feedforward and the feedback neural network (backpropagation). Examples of

feedforward neural networks include multilayer perceptron (MLP), probabilistic neural network, Adaline and Madaline neural network while examples of backpropagation neural networks include Hopfield network, bi-directional and associative memory (BAM) neural network [15, 16]. Several ANN techniques have been used for different applications e.g. prediction of weather information, optimization of power in solar energy systems, classification of data samples, etc. [17]. This paper will focus on three-known ANN training techniques (Levenberg-Marquardt, scaled conjugate gradient, and Bayesian regularization ANN algorithms).

The Levenberg-Marquardt (LM) is a hybrid technique that combines the fast converging speed of Gauss-Newton algorithm and the stable capability of the steepest descent method for training process into a single algorithm [18]. The mathematical model of LM algorithm can be defined using equation (1) while the dual working operation of Levenberg-Marquardt algorithm is explained using equations (2) and (3),

$$w_{k+1} = w_k - (J_k^T J_k + \mu I)^{-1} J_k e_k \quad (1)$$

$$w_{k+1} = w_k - (J_k^T J_k)^{-1} J_k e_k \quad (2)$$

$$w_{k+1} = w_k - \alpha g_k \quad (3)$$

where μ is the combination coefficient, *I* is the identity matrix, J is the Jacobian matrix, $J^T J$ is the Hessian matrix, g is the gradient vector, e is the error vector, α is the learning constant (step size), w is the weight vector and k is the index of iteration [19]. When μ is very small (approaching zero), the Levenberg-Marquardt switches from equation (1) to equation (2) known as Gauss-Newton algorithm. Similarly, when μ is very large, the descent method represented by equation (2) is used [20, 21].

The Bayesian regularization is another ANN algorithm introduced lately and have been used by research scholars to solve real-world problems. The Bayesian regularization helps to reduce noise in the training data of ANN and ensures that smoother network-response. The Bayesian algorithm helps to accommodate large weight vector in ANN by fine-tuning the used objective function and with the addition of a penalty term that comprises of squares of all network weights [22, 23].

The scaled conjugate is an algorithm commonly used to train networks that have a large number of weights due to the rapid training speed that the algorithm exhibit [24, 25, 26, 27].

III. SIMULATION MODEL

To examine the feasibility of the proposed noise mitigation technique modelled using different ANN algorithm to identify the most suitable algorithm that can effectively remove the impulsive noise (IN) existing in the transmitted signal (T_x) in an OFDM channel, a case study was done with three different ANN techniques (Bayesian, Levenberg Marquardt, and Scaled cojugate gradient).

A complete OFDM system that comprises of Bernoulli binary generator for generating the transmitted bit signals of 0 and 1 (T_x), binary phase shift keying (BPSK) modulator, Inverse fast fourier transform (IFFT), FFT (fast fourier transform), BPSK demodulator, additive white Gaussian noise channel and randomly-corrupt impulsive-noise data of variance of 50% and probability of 50% were added to the OFDM model to achieve the objective of the experiment. The random impulsive noise datasets were produced using equations (4-6),

$$F_m(n_k) = \sum_{m=0}^{\infty} p_m \mathbb{N}(n_k; 0, \sigma_m^2) \quad (4)$$

$$\text{where,} \quad P_m = \frac{A^m e^{-A}}{m!} \quad (5)$$

$$\text{and} \quad \sigma_m^2 = \sigma_i^2 \frac{m}{A} + \sigma_g^2 = \sigma_g^2 \left(\frac{m}{A\Gamma} + 1 \right) \quad (6)$$

$F_m(n_k)$ is the probability density distribution (PDF) of a noisy signal (n_k), $\mathbb{N}(n_k; \mu, \sigma_m^2)$ expresses the Gaussian PDF, μ is the mean, σ^2 is the variance with k -samples. σ_i^2 denotes the impulsive noise variance, σ_g^2 is the additive white Gaussian noise (AWGN). $\Gamma = \frac{\sigma_g^2}{\sigma_i^2}$ denotes the Gaussian-to-impulsive-noise power ratio. The parameter (A) is the density of impulses within a specific width and observation period.

Figure 2 displays the block diagram of a complete OFDM system modelled using ANN error-correcting schemes. The ANN training was conducted using Bayesian-regularization, Levenberg-Marquardt, and Scaled-conjugate gradient algorithm. These supervised ANN controllers learn using 10,000 samples of signals u and v, where u is the transmitted signal (T_x) mixed with impulsive noise and AWGN, and v-the impulsive noise are the input variables ***(u,v)*** (predictors) and signal t (output signal from the IFFT/AWGN block) is the target. For simplicity, the complex-variable data $u = u_1 + ju_2, v = v_1 + jv_2,$ and $\vec{t} = \vec{t}_1 + j\vec{t}_2$ were split into real and imaginary components and both the real and imaginary parts of signals (u, v) were the ANN inputs (u_1, u_2, v_1, v_2) while the real and imaginary component of signal t (t_1, t_2) were used as targets. The ANN channel is then applied just before the BPSK demodulation in order to mitigate the impulsive noise from the trasmitted data.

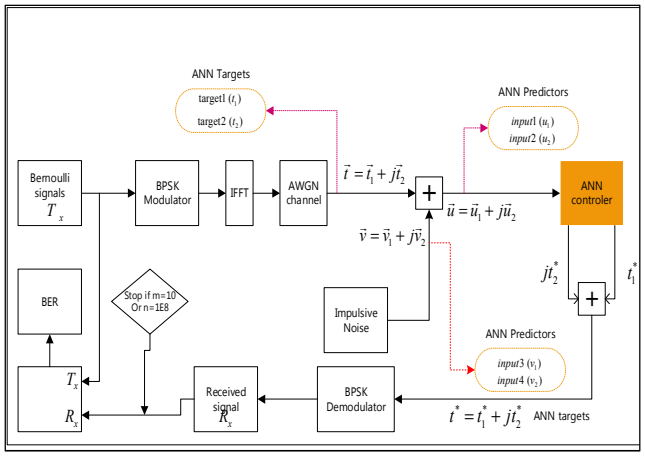


Fig. 2: Complete OFDM channel Using ANN Techniques

Figure 3 presents the amplitude level of the corrupted impulsive noise signal in the OFDM channel for 500 seconds.

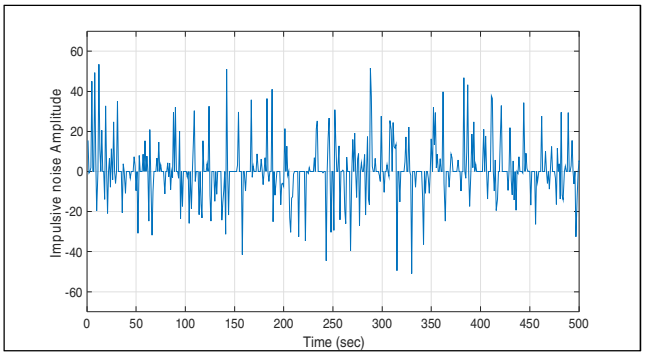


Fig. 3: Impulsive signal Magnitude

IV. EXPERIMENTAL RESULTS

Figures (4) – (7) and Tables (1) - (2) present the graphical results and tabulated results of the used ANN error correcting techniques done using MATLAB simulation approach respectively. From Table I, it can be seen that Bayesian regularization technique exhibited the best training performance with the lowest mean square error (MSE is 6.1834e-11) while scaled conjugate gradient exhibited the worst performance (MSE is 2.7545e-2). However, Bayesian regularization takes a longer time to train as the training for 1000 epochs was completed in 7 mins 20 seconds while Levenberg-Marquardt had the fastest training time (32 seconds).

Table 1: Parameters for the three ANN algorithms

	Bayesian Regularization	Levenburg Marquardt	Scaled Conjugate Gradint
Hidden neurons	10	10	10
Training MSE	6.1834e-11	5.6757e-10	2.7545e-2
Validation MSE	0	5.7723e-10	2.8855e-2
Testing MSE	6.9677e-11	6.9364e-10	2.6541e-2
Traing Regression	9.9999e-1	9.9999e-1	9.7451e-1
Validation Regression	0	9.9999e-1	9.7348e-1
Testing Regression	9.9999e-1	9.9999e-1	9.7519e-1
Epoch	1000	1000	124
Perfomance	6.18e-11	5.68e-10	0.0273
Time	7mins 20secs	32secs	33secs
Gradient	2.07e-6	1.27e-5	0.0135

Figure 4 displays the bit error rate (BER) graphical results of a comparison done with Bayesian regularization ANN impulsive-noise error-correcting technique and conventional (uncorrected) method using an OFDM system that has been corrupted with both both AWGN and impulsive noise. From the obtained graphical results, the high performance of the Bayesian regularization technique can be observed as it requires less than 10 dB power to achieve a BER of 10^{-4} whereas with an uncorrected OFDM channel requires a high SNR power.

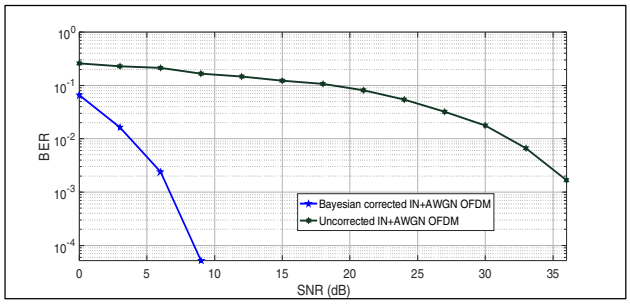


Fig 4: BER performanne of Bayesian Regularization algorithm

Figure 5 shows the BER results of the OFDM chanel (mixed with AWGN and IN) corrected using levenberg marquardt error correcting algorithm compared with the conventional uncorrected OFDM system. From the obtained results, the high impulsive-noise mitigating capability of the Levenberg-Marquardt can be seen as it requires less than 10 dB to achieve a BER of 10^{-4}, while the uncorrected corrupted OFDM channel consumes more SNR power.

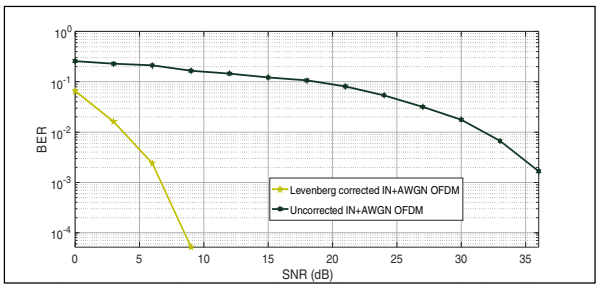


Fig 5: BER performance of Levenberg-Marquardt algorithm

Figure 6 presents the bit error rate (BER) performance of the OFDM channel (mixed with impulsive noise and AWGN) incorporated using scaled conjugate gradient ANN algorithm compared with the uncorrected OFDM system. From the obtained results, it can be seen that using scaled conjugate method requires approximately 10 dB SNR power to achieve a BER of 10^{-4}.

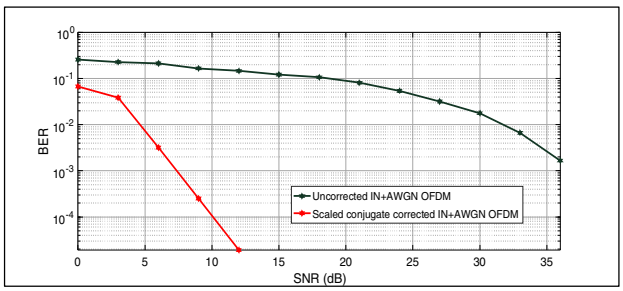


Fig 6: BER performance of scaled Conjugate Gradient

Figure 7 shows a comparison of the BER performances of the above-mentioned ANN algorithms (Levenberg, Bayesian, and scaled-conjugate) results and the conventional un-corrected OFDM results. From the obtained results, it can be seen that both Levenberg-Marquardt and Bayesian regularization requires the lowest SNR power for the mitigation of impulsive noise in a corrupt OFDM channel.

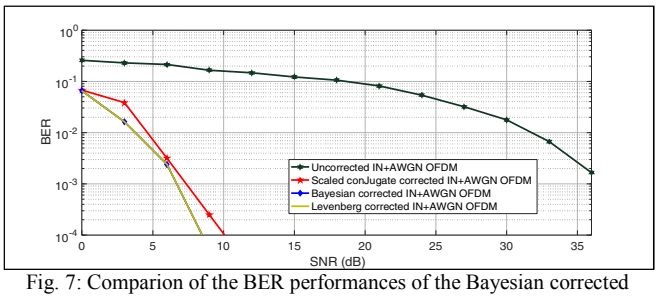


Fig. 7: Comparion of the BER performances of the Bayesian corrected OFDM, Levenberg-Maquardt corrected OFDM, scaled-conjugate corrected OFDM and the uncorrected OFDM

Table 2 displays the tabulated bit error rate (BER) results under different SNR (signal-to-noise ratio) power in decibels. T is the theorectical results for OFDM channel mixed with AWGN only. SC is the scaled conjugate error correcting algorithm, BR is the Bayesian regularization, LM is the Levenberg-Marquardt, and CO is the convential uncorrected results for OFDM channel that was corrupted with both impulsive noise and AWGN respectively.

Table 2: BER results for the theoretical, scaled-conjugate, Bayesian regularization, Levenberg-Marquart algorithm, and conventional uncorrected OFDM channel under different SNR power

SNR	T BER	SC BER	BR BER	LM BER	CO BER
0	0.0786	0.067110	0.06579	0.06579	0.2577
3	0.0228	0.038460	0.01623	0.01623	0.2278
6	0.0023	0.003177	0.00239	0.00239	0.2119
9	3.36e-5	2.500e-4	5.177e-5	5.177e-5	0.1658
12	9.00e-9	1.918e-5	-	-	0.1460
15	9.1e-16	-	-	-	0.1220
18	1.4e-29	-	-	-	0.1063
21	5.3e-57	-	-	-	0.08091
24	1e-111	-	-	-	0.0540
27	2e-220	-	-	-	0.03185
30	-	-	-	-	0.01768
33	-	-	-	-	0.006662
36	-	-	-	-	0.001662

From the obtained results (see Fig. 4-7 and Tables 1-2), it was observed that both Bayesian and Levenberg-Marquardt exhibited an effective impulsive noise error correcting capability while scaled conjugate algorithm displayed the lowest impulsive noise error-correcting performance.

V. CONCLUSIONS

This paper presents an innovative use of Levenberg-Marquardt and Bayesian regularization ANN machine learning techniques for the improved mitigation of impulsive noise in an Orthogonal Frequency Division Multiplexer (OFDM) channel. To validate the efficiency of the above-mentioned ANN techniques (Levenberg-Marquadt and Bayesian regularization), an OFDM channel that lacks an error-correcting scheme simulation results was compared with Levenberg-Marquardt results, Bayesian regularization results and with the results of another popular ANN technique (scaled conjugate) in order to validate the importance of impulsive noise error-correcting scheme and thus evaluate the effectiveness of the used error-correcting methods. Findings suggest that both Levenberg-Marquardt and Bayesian regularization exhibit better performance in removing the impulsive noise in the OFDM channel and required minimal signal-to-noise (SNR) power.

REFERENCES

[1] J. A. Cortes, L. Diez, and F. J. Canete, "Analysis of the periodic impulsive noise asynchronous with the mains in indoor PLC channels," in IEEE International Symposium on Power Line Communications and Its Applications, Dresden, Germany, 2009.

[2] C. Cano, A. Pittolo, D. Malone, L. Lampe, A. M. Tonello, and A. G. Dabak, "State of the art in power line communications: From the applications to the medium," IEEE Journal on Selected Areas in Communications, vol. 34, no. 7, pp. 1935-1952, 2016.

[3] D. B. Ünsal and T. Yalçınöz, "Applications of new power line communication model for smart grids," International Journal of Computer and Electrical Engineering, vol. 7, no. 3, pp. 168-178, 2015.

[4] L. Bai, M. Tucci, S. Barmada, M. Raugi, and T. Zheng, "Impulsive Noise Characterization in Narrowband Power Line Communication," Energies, vol. 11, no. 4, pp. 863-880, 2018.

[5] N. M. Manikanda and M. Chandrasekaran, "An efficient and low power impulsive noise suppressor architecture for OFDM system," Cluster Computing (Springerlink), vol. 2018, pp. 1-7, 2018.

[6] J. Radić and N. Rožić, "Adaptive impulse noise suppression in OFDM systems," Elektronika, vol. 1, no. 94, pp. 1-6, 2010.

[7] T. Shongwe, A. J. Han Vinck, and H. C. Ferreira, "On Impulse Noise and its Models," in 18th IEEE International Symposium on Power Line Communication and its Applications (ISPLC), Glasgow, UK, 2014.

[8] X. Zhang, Z. Yuping, and L. Zou, "An efficient impulsive noise mitigation scheme for over-sampled OFDM systems," IEEE Transactions on Consumer Electronics, vol. 55, no. 2, pp. 360-365, 2009.

[9] S. V. Zhidkov, "Impulsive noise suppression in OFDM-based communication systems," IEEE Transactions on Consumer Electronics, vol. 49, no. 4, pp. 944-948, 2003.

[10] M. R. Ahadiat, P. Azmi, and A. Haghbin, "Impulsive noise estimation and suppression in OFDM systems over in-home power line channels," International Journal of Communication Systems, vol. 2014, no. 30, pp. 1-15, 2014.

[11] K. Ogura, A. Nakamura, and M. Itami, "A study on class A impulse noise cancellation and channel estimation under Rayleigh fading environment," in IEEE International Conference on Consumer Electronics (ICCE), Hue City, Vietnam, 2018.

[12] S. Deepa and S. Kuzhaloli, "A survey on interleavers used for OFDM systems on various applications," International Journal of Pure and Applied Mathematics (IJPAM), vol. 118, no. 17, pp. 499-507, 2018.

[13] T. Shongwe, A. J. Han Vinck, and H. C. Ferreira, "The effects of periodic impulsive noise on OFDM," in 2015 International Symposium on Power Line Communications and its Applications (ISPLC), Austin, TX, USA, 2015.

[14] R. Chol Ho, S. Eisenbarth, and Y. Chon Ri, "Performance evaluation of OFDM transmission over impulsive noisy channels," in 2017 IEEE International Conference on Computer, Information and Telecommunication Systems, Dalian, China, 2017.

[15] J. Zhang, H. Zhiqiang, P. Chen, and Y. Rong, "A compressive sensing Based iterative algorithm for channel and impulsive noise estimation in underwater acoustic OFDM systems," in OCEANS 2017 - Anchorage, Anchorage, AK, USA, 2017.

[16] A. Al-Dweik, A. Hazmi, B. Sharif, and T. Charalampos, "Efficient Interleaving Technique for OFDM System over Impulsive Noise Channels," in 28th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Istanbul, Turkey, 2010.

[17] Y. Himeur and A. Boukabou, "OFDM-based power-line communication enhancement using a turbo coded adaptive impulsive noise compensator," Telecommunication Systems (Springerlink), vol. 2015, pp. 1-14, 2015.

[18] C. Anusuya, M. Jothimani, and R.Kalaivani, "OFDM based power line communication intensification using RSC coding with adaptive noise compensator," in International Conference on Emerging trends in Engineering, Science and Sustainable Technology (ICETSST-2017), Erode, India, 2017.

[19] R. Barazideh, A. V. Nikitin, and B. Natarajan, "Practical implementation of adaptive analog nonlinear filtering for impulsive noise mitigation," in IEEE International Conference on Communications (ICC), Kansas City, Kansas, USA, 2018.

[20] A. N. Hasan and T. Shongwe, "Proposed machine learning system to predict and estimate impulse noise in OFDM communication system," in 42nd Annual Conference of the IEEE Industrial Electronics Society, IECON 2016, Florence, Italy, 2016.

[21] P. Chen and Y. Rong, "Joint channel estimation and impulsive noise mitigation in underwater coustic OFDM communication systems," IEEE Transactions on Wireless Communications, vol. 16, no. 9, pp. 6167-6178, 2017.

[22] A. N. Hasan and T. Shongwe, "Impulse noise detection in OFDM communication system using machine learning ensemble algorithm," in Advances in Intelligent Systems and Computing, San Sebastian, Spain, Springerlink, 2017, pp. 1-10.
[23] A. M. Farayola, Comparative study of different photovoltaic MPPT techniques under various weather conditions (Thesis submitted in partial fulfillment for the degree of Masters in Electrical Engineering), Johannesburg, South Africa: University of Johannesburg, 2017.
[24] A. M. Farayola, A. N. Hasan, A. Ali, and B. Twala, "Distributive MPPT approach using ANFIS and perturb&observe techniques under uniform and partial shading conditions," in International Conference on Artificial Intelligence and Evolutionary Computations in Engineering Systems & Power, Circuit and Information Technologies (ICPCIT-2017), India, Springer, Singapore, 2018, pp. 27-37.
[25] A. M. Farayola, A. N. Hasan, and A. Ali, "Curve fitting polynomial technique compared to ANFIS technique for Maximum Power Point Tracking," in 8th IEEE International Renewable Energy Congress (IREC), Amman, Jordan, 2017.
[26] A. M.Farayola, A. N. Hasan, and A. Ali, "Optimization of PV systems using data mining and regression learner MPPT techniques," Indonesian Journal of Electrical Engineering and Computer Science (IJEECS), vol. 10, no. 3, pp. 1-9, 2018.
[27] M. Zimmermann and K. Dosert, "Analysis and modelling of impulsive noise in broadband powerline communication and visible light communication," IEEE Transaction on Electromagnetic, vol. 44, pp. 249-258, 2002.

The Value Perspective to Deploying a Cloudlet Business Model for SMEs

Mandisa N. Nxumalo, Ijeoma N. Mba and Matthew O. Adigun
Department of Computer Science,
University of Zululand
South Africa
mandisan.mn@gmail.com, ijaymba@gmail.com, profmatthewo@gmail.com

***Abstract*—The provision of Cloud services through the use of a Cloudlet technology, provides a number of advantages with regards to Quality of Experience (QoE) for its consumers. The QoE includes, but is not limited to free of cost, low-latency, and one-hop WiFi network consumption. Most of the advantages are achieved through the deployment of a comprehensive business model. A well-defined business model is one with well-defined value logic. Value creation is a fundamental element of a business model. It helps identify relevant customer segments, value proposition for customers and mechanism that will be used to provide the created value. However, poorly defined value logic into a business model can lead to business failure in addressing their customers' needs and issues. The literature reveals that researchers delved more into the evaluation of Cloud business models and its subsequent value creation, but none has attempted on either creation or deployment of a Cloudlet Business model. Therefore, this study introduces a Cloudlet Business Model (CBM) that can be used to support the easy deployment of a Cloudlet technology by SMEs such as coffee shops and shopping malls. In order for SMEs to meet their customers' needs and be able to sustain their businesses. This study undertakes the use of a Four Box model to bring out the value perspective of a CBM. This resulted in the identification of the value as small data, how it is generated, delivered and compensated for especially cost in ensuring the effectiveness of a CBM on the SMEs.**

Keywordst*—*Cloudlet, QoE, SLA, small data, value logic

I. INTRODUCTION

The value proposition is an important aspect of a business model. It gives details on, why a business would intend to deploy certain services and how all stakeholders involved will benefit from it. A business model in this context refers to the value logic of a business in terms of how it creates and capture customer's value [1]-[3]. An effective business model is said to consist of an interrelated set of elements that address the customer value proposition, business architecture (Key Resources and Processes) and the economic dimensions (Profit Formula) [4], [5]. These elements are parameters of strategic management techniques used to develop new or document existing business models. Examples of strategic management techniques include the Canvas Business Model, Four Box Business Model, and others [6]. These techniques are responsible for clarifying value logic of deploying a certain business model in an enterprise.

Small, Medium Enterprises (SMEs) consider the transitioning to the Cloud very risky due to the lack of hardware resources and a high cost of leasing its resources such as internet bandwidth to avoid WAN latency [7]. An example of SMEs in discussions are businesses such as a coffee shop, internet café, doctor's office, shopping malls, campuses, airports or libraries [8]. For such businesses, deploying a Cloudlet on their premises can be of benefits. The benefits include low deployment cost when compared to Cloud which will ease the deployment. This is because the deployment of a Cloudlet allows the reuse or investment in economical hardware's such as wireless routers. Also, due to the present state of accessing Cloudlet resources through WLAN (Wireless, Local Area Network), a business can seize the opportunity to improve the purchase patterns of their core service or product being offered. Meaning deploying a fast, low latency, non-monetary, high bandwidth technology for internet access can attract customers to stay, which can lead to more consumption of the offered services. This will surely improve the marketing, review ratings, meeting customers' QoE and profit for the business.

A Cloudlet is described as a data center in a box, introduced by Mobile Cloud Computing (MCC) to eliminate Wide Area Network (WAN) latencies encountered by Cloud consumers when accessing response sensitive applications [8]-[11]. Examples of response sensitive applications are not limited to the use of augmented reality, face recognition, online gaming and video streaming [9]. A Cloudlet is more efficient for the above stated applications as compared to Cloud due to its proximity by the edge (near mobile users). As shown in Fig.1, it is a one-hop, resource-rich computer or cluster of computers that allows mobile users to consume Cloud resources and services without encountering high latency [11]. It uses 3-tier architecture (Mobile users – Cloudlet - Cloud). The Cloud offloads resources to the Cloudlet and the Cloudlet uses a one hop connection to provision the resources to end users. However, despite overcoming latency issues the deployment of a Cloudlet by SMEs remains a challenge [7], [12]. This is due to the lack of information about its business models to help facilitate its easy deployment by SMEs. The lack of Cloudlet business models remains a challenge and a risk for the previously mentioned SMEs to trust that its deployment on their premises can yield to positive outcomes.

This paper pays a close attention to the significance a Cloudlet deployment can have on the business operations and its customers. Mainly focusing on the logic value of using a CBM, to which we intend to answer questions such

as what the actual value of CBM is and what benefits can be seized from that value to benefit the parties involved. Furthermore, discuss on mechanisms that can be used to capture the value and compensate it. The Four Box Model was used as a strategic management technique to discuss a Cloudlet Business Model (CBM), shown in Fig.4. This model helps provide clarity on how the deployment of CBM by SMEs can create, capture value and yield benefits for all parties involved.

This paper is organised as follows: Section 2 introduces three key concepts (the stakeholders, SLA, and small data) used in the CBM. Section 3 discusses related work to this study. The value logic of CBM derived from the use of the Four Box Framework is discussed in Section 4. Lastly, Section 5 provides the conclusion of the paper and future work.

II. The Concept of Stakeholders, SLA and Small Data in the CBM

The CBM consist of several components (shown in Fig.2) that communicate with each other to ensure a successful execution of a service based on an agreed SLA by stakeholders. The aim of the CBM is to facilitate the easy deployment of a Cloudlet infrastructure in business premises [12]. Therefore, there is a need to identify the stakeholders involved in the CBM. This model consists of three main stakeholders or role players:

- Cloudlet Consumers (CC): are Cloudlet resource consumers willing to share their browsing history (small data) in order to consume Cloudlet resources.
- Cloudlet Owner (CO): An individual or group of individuals (SMEs) that deploys a Cloudlet in their business premises with the aim to attract consumers and increase the purchase value of their core product/service.
- Cloud Service Provider (CP): are companies such as Google and Microsoft that sell their resources to CO using an agreed price function.

The deployment and the relationship between the mentioned role-players are facilitated by a standard protocol called a Service Level Agreement (SLA) that describes the offered service, its guarantees, and penalties involved [12]-[15]. The SLA plays a significant role in providing clarity of the business value logic, as it is used to negotiate the terms involved between entities [12]-[15]. The CBM consists of two SLAs, namely SLA1 (agreement between CC and CO) and SLA2 (agreement between CO and CP) as shown in Fig.3. SL1 is issued when a mobile user sends a request for processing to a Cloudlet and is received in a form of a captive portal. A captive portal is used to enforce the Cloudlet access policies to the mobile users by enabling them to agree in sharing their browsing history for Cloudlet access. Where else, SLA2 is issued when a CO intends to purchase services from CP or Internet Service Provider (ISP) for enabling a high QoE internet service. The existence of the two SLA's is to develop a model that ensures fairness to all role-players for a successful deployment of a Cloudlet by the SMEs. Although both SLA's have a high impact on ensuring service guarantees but this study pays more attention to SLA1. The SLA1 is measured using metrics such as availability, CPU utilization, cost, waiting time, response time (or task completion time) and fault tolerance level. The success of meeting the above stated SLA1 metrics contributes highly on a successful CBM deployment and execution of a business value logic, where else violations can lead to a business downfall. This is because SLA1 violations can lead to penalties that might be costly for CO and can also deprive customers to come or stay on the CO's business premises. To understand and reflect the function of the SLA1 and how it is captured will be carried out using a simulation tool called CloudSim Plus [16]. Assuming that the SLA1 negotiations occur autonomously and customers hav agreed to sharing their browsing history, the simulation will focus on assessing if the SLA1 metrics are met starting when the CC issues a request for computation to the Cloudlet, up until the job is delivered into the CC. Fig.2 shows a function that will be used to compare the stated task completion time on SLA1 (shown in Fig. 4) with the actual task completion time taken by the Cloudlet during the simulation. The SLA1 contract is written in javascript and is fed into an SLA contract class which captures the dimensions of the SLA1 to make it easy to compare the stated SLA1 parameters with obtained simulation results. This simulation is done to show how the SLA component will work to ensure service guarantees on the CBM.

One of the CBM objectives is to improve customer's quality of experience when consuming Cloudlet resources, through the usage of small data. Small data refers to the sharing of context information that describes a mobile user's browsing patterns, clearly pointing out the interests and preferences of a mobile user. Data is described as small data if it is accessible, informative and actionable [22], [23]. This data is obtained through a customer's consent. Before a customer can consume Cloudlet resources they need to agree to terms stated on the captive portal. The captive portal aim is to get approval for a Cloudlet to get access and capture the mobiles user's browsing history (small data). The small data patterns are obtained or observed from a log file generated when a mobile user uses mobile browsing software to surf the net. The log file consists of a history of sites a mobile user visited and through this a Cloudlet can recognize the mobile users browsing patterns to predict future access.

Data stored on the Cloudlet cache is retrieved or prefetched from the Cloud when CC sends a similar request for processing. For an example in cases where a mobile user request for an already existing service, the Cloudlet will prefetch the service from its cache to the CC, minimizing the response time, CPU utilization and energy consumption. The caching of browsing history is already achieved by one of the Cloudlets commercial products called the Xiaomi Smart Router [24]. The router allows web prefetching, video prefetching and file backup. Web prefetching allows the capturing of small data through the use of a Sougou prefetching engine that enables the router to make predictions of CC preferences using their browsing history (historical access logs). The capturing of small data in the CBM is considered as the cost of consuming Cloudlet resources by the mobile users. In cases where the small data is not captured the request will be directed to the Cloud, assuming that the SLA1 was rejected.

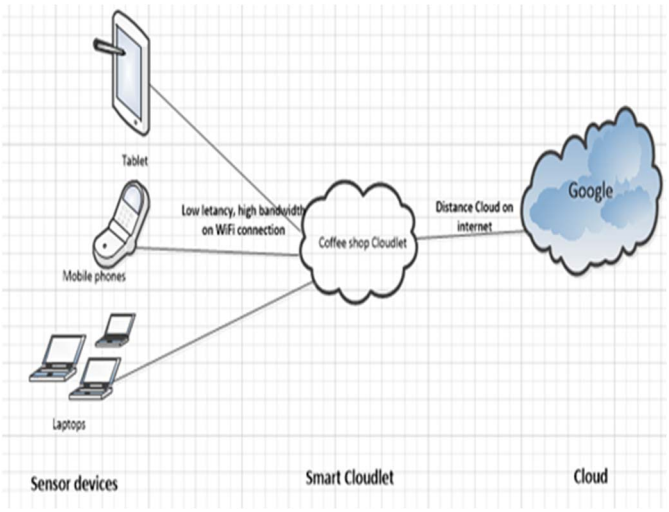


Fig. 1. Cloudlet Architecture

```
{
  "metrics": [
    {
      "name": "TaskCompletionTime",
      "dimensions": [
        {
          "name": "minValue",
          "value": 0,
          "unit": "Milliseconds"
        },
        {
          "name": "maxValue",
          "value": 100,
          "unit": "Milliseconds"
        }
      ]
    },
```

Fig. 4. Sample of SLA1 contract with Task Completion time metric

```
double getPercentageOfCloudletsMeetingCompletionTime() {
    final DatacenterBroker broker = getFirstBroker();

    final double totalOfcloudletSlaSatisfied = broker.getCloudletFinishedList().stream()
            .map(c -> c.getActualCpuTime() + c.getWaitingTime())
            .filter(rt -> rt <= getMaxTaskCompletionTime(broker))
            .count();

    final double percent = totalOfcloudletSlaSatisfied * 100 / broker.getCloudletFinishedList().size();
    System.out.printf("\n # Total of cloudlets satisfied SLA completion time of %.2f secs: %.0f of %d",
        getMaxTaskCompletionTime(broker), totalOfcloudletSlaSatisfied, broker.getCloudletFinishedList().size());
    System.out.printf("\n # Percentage of cloudlets that complied with the SLA Agreement:  %.2f %%", percent);

    System.out.println("\n\nVirtual Machines: " + broker.getVmCreatedList().size());

    return percent;
```

Fig. 2. Method for verifying the response time SLA metric

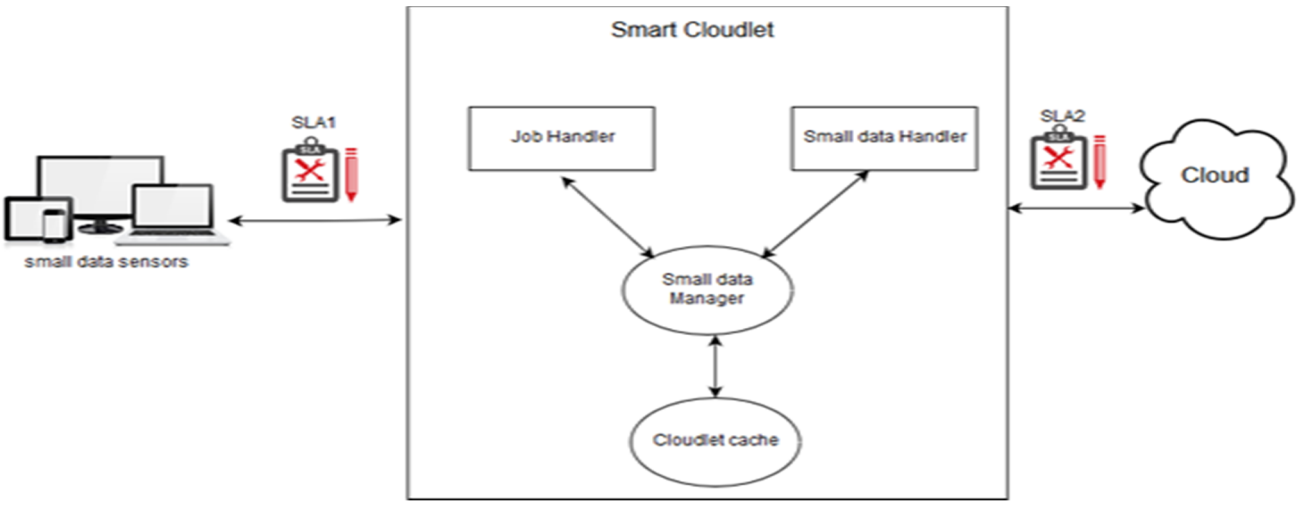


Fig. 3. An envisaged Cloudlet Business Model

III. Literature-Review

The value proposition seems to be a trending word for business model development and its improvement [1]-[6]. It acts as an effective merger to take informative decisions on whether to adopt a certain business model or not. Also, as an indicator of a business model success rate [6]. In literature not much information is available concerning the deployment of a Cloudlet technology and the value it can create for its owners and consumers. The below provided literature form state of art in ensuring an effective business model development that meets customers' needs and yield benefits for its owners. The contribution by other authors will provide guidance on how a CBM can ensure that its value proposition is clear and can yield to ensuring service guarantees for the CC.

An article covered by [4] pays close attention on the difference between a Four Box Model and Canvas Business Model. A Canvas Business Model is visual chart that consists of elements describing a business or product's value proposition, infrastructure, customers and finances. It can be used to brainstorm concepts of a business model and visualise their interactivity. The Four Box Model was introduced by Mark W. Johnson to help businesses innovate their business models into white-spaces that are highly profitable and less competitive. The framework focuses on the interaction between a customer value proposition, profit formula, key resources and processes to interpret a business model. Based on the comparison the Four Box model is described as more effective than Canvas due to its simplicity in highlighting the interactions between the elements more greatly [21].

In ensuring a sustainable business model design, some scholars introduced a value mapping tool [17]. The value mapping tool takes into consideration the different forms of value exchanges for a range of stakeholders as part of the business model development. The different forms of values discussed in this literature are the value captured, value missed, value destroyed and new value opportunities. The applicability of the tool was tested in a workshop setting and the findings indicated that the tool can be used to explore business sustainability issues, clarify needs and expectations of stakeholders, and can be used as an ideation for Non-Government Organisations (NGO's).

A theory and practice understanding of value creation and its purpose on a business model was discussed in [19]. The value is described as an entity created through an organization's business model. That takes into account inputs from the capitals and transforms them through business activities and interactions. In order to produce outputs and outcomes that over time will create or destroy value for the organization, its stakeholders, society, and the environment. The drawback of this work is that it does not cover how the value is compensated in the case studies presented.

A Value Network Development Approach was introduced to help in bringing out the business value logic [20]. This approach was used to answer the effect of using a Canvas Business Model Canvas as a methodology for value creation. The authors followed the five steps in their approach to value creation: (1) identification of suitable collaboration and governance mechanism for value networks, (2) identification of role-players in value networks and their roles, (3) clarification of advantages and disadvantages for network value role-players, (4) identification of exchanges and benefits involved for role-players. Lastly, (5) identification of GRID opportunities for role-players to align their business model. Although, the Business Model Canvas is easy to change, visualize and understand. However, it lacks a more detailed operational and financial aspect when compared to a Four Box Framework.

A qualitative study (distribution of questionnaires to collect data for analysis) for evaluating the efficiency of the created value of a Cloud service by operational businesses was conducted in [18]. This was done to gain an understanding of how Cloud services are used in different organisations to improve their operations. Based on the information gathered, the authors then proposed a Cloud service governance approach to fully address the concept of value importance in a business model. Some of the suggestions imposed on this literature with regards to existing businesses are the improvement of a business model to ensure that it addresses previously unsolved customer concerns and needs.

The CBM uses the Four Box models to clearly describe the interaction between different components of the model and the SLA to ensure the value logic is executed to ensure service guarantees for CC.

IV. Value Logic of CBM

The value logic of CBM is explained through the use of a Four Box model [4], [6]. This model consists of four elements, namely Customer Value Proposition, Profit Formula, Key Resources and Key Processes. The discussion below is based on how the above-mentioned elements (shown in Fig. 5) are applied in CBM to show interaction between CBM components.

A. Customer Value Proposition (CVP)

The CVP element requires the identification of a problem a customer is encountering that a product or service proposed will optimize [4], [19]. In accordance to CMB, the CVP is an offering that helps customers to effectively, reliably, conveniently, or affordably solve a poor Quality of Experience (QoE) problem in the usage of Cloud services at the expense of small data sharing. The small data is analysed by the Small data handler component which passes it through the Small data manager to check for its existence on the cache. This is done to avoid data redundancy. The caching of browsing history is done with the aim to provide CC with services of preference in a reduced execution time and processing energy that might have been used to process the request. The known fact about Cloudlet-WiFi is that CC doesn't have to pay a direct cost for its resource consumption [13]. As mentioned earlier that small data can be used to bill Cloudlet resource consumption, in this way resources will be managed, processing rate decreased, and SLA objectives met. It is for this reason we believe that deploying a resource rich, one hop, non-monetary, low-latency technology can be valuable for customers and can improve their surfing experience.

A Coffee Shop Use Case: Consider a coffee shop business that is currently using or have deployed a Cloud-WiFi service for its customers to consume internet services. An assumption is that the customers will be consuming the Cloud services for free, which means that the Coffee shop owners are the ones paying for the consumption. They benefit from the deployment through increasing the core product price, in order to adapt to high energy costs resulting from the deployment. This cause of action is sometimes not effective towards improving customer QoE. Also, it can increase the probability of customers who use the Cloud-WiFi service, but not purchase anything from the coffee shop. This can have a negative effect on the business. If the business was to deploy a Cloudlet-WiFi due to its merits the business can improve its customer QoE when consuming Cloud service and the purchase and marketing strategy of its core product. Assuming that customers will spread the word about their experience of using a one-hop, low-latency, non-monetary, resource rich technology deployment on a coffee shop. The increase in customer interest to spend time at the Coffee Shop due to the QoE obtained from using a Cloudlet can result to an increase in business profit. Meaning customers might end up consuming Coffee Shop core services, improving core product sales.

B. *Profit Formula*

The Profit Formula defines how the business will create value for itself and its customers. It specifies the revenue model, cost structure, target unit margins and how quickly resources need to be used to support a target volume [4], [19]. This particular element answers the question related to the strategy that will be used to capture the roles of CBM players. This will help generate revenue for the CO and provide fairness for all parties involved. The CBM uses two price functions as its strategy due to the nature of the service and different stakeholder's role. For instance, the CC since they will share context to consume Cloudlet resources, they will be using a *pay less for bulk usage* price function. However, the threshold of bulkiness is controlled by the scheduling policies implemented on a Cloudlet job handler component to manage its resources. On the other hand, the CO will use a *pay for what you use* price function to pay for the consumption of Cloud services by its customers. Both price functions are captured on the SLA.

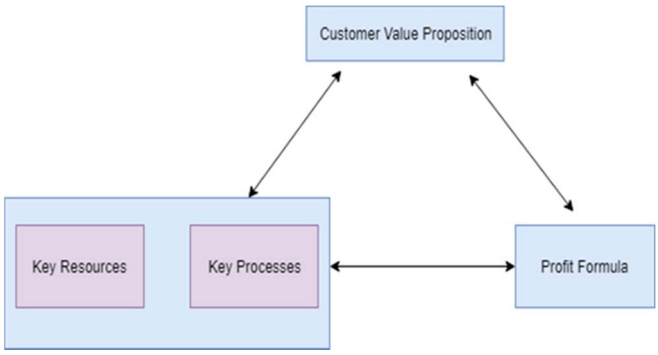


Fig. 5. A Four Box Model [4]

C. *Key Resources and Processes*

The two elements of technique namely, Key Processes and Resources required for the CBM to deliver the value logic are discussed below:

1) Key Resources: this element describes the entities (people, technology, equipment, information, partnership) required to deliver the value proposition to the customers [4], [19]. In the CBM the required entities are Cloud providers or Internet Service Providers (ISP); equipment and technology companies that will provide tools such as a router and also help install the devices. The Key resources of the CBM cover all its components and role-players.

2) Key Processes: describes the mechanism and policies a business implements to facilitate the successful delivery of a value proposition to the customers [4], [19]. The CMB uses an SLA as a standard protocol for facilitating the deployment of the service and the delivery of CVP. In CBM, when the CC initiates a request the SLA executor on the Job handler component is initiated to check if the CC has given a Cloudlet the rights to capture their browsing history. Another mechanism incorporated in the CBM is the task scheduling policies (Space-based or time-shared scheduling policy) to control task processing and resource management on the Cloudlet environment.

The elements discussed above have a direct impact on each other. In order to define a billing method, there has to exist a service that will bring significance to your customers. Furthermore, to maintain the service, its processes and ensure fairness amongst involved parties requires placement of standards or policies to facilitate the delivery of that service. Lastly, the deployment of the service requires certain resource collaboration and equipment in place. In relation to CBM, these elements help in bringing out detailed operational information (in relation to business rules, responsibilities of stakeholders and success metrics) and financial features important for clarifying the value logic of CBM.

V. CONCLUSION

The main goal of the CBM is to ensure that the created value solves SME customer issues and meets the needs of both customers and SME stakeholders. Based on the provided discussion, the benefits that can be enjoyed by CC are not limited to free of cost, low latency, one-hop Wi-Fi network consumption based on their preferences. Furthermore, the deployment of a Cloudlet in SME environments can attract more customers which can lead to an increase in buying patterns of the offered product by SMEs. In this paper, the value perception of a CBM was discussed. The discussion was carried out through the use of a Four Box Model. Part of the contribution proposed in this paper was the use of small data (value of CBM) as means of compensating the consumption of Cloudlet resources to provide services of preference to CC (improve QoE), manage Cloudlet resources and to reduce energy costs. Also, the usage of an SLA to capture the business value logic, its processes, economic dimensions and penalties involved in cases of failure was discussed. It can be concluded that the success of a business model lies in its ability to clearly paint out its value logic and the benefits it will bring to all parties involved. Further work involves the simulating of how the SLA processes using CloudSim Plus an extension of CloudSim. Also, the usage of the Cost-benefit Analysis to prove the feasibility of deploying a Cloudlet covering the costs and benefits involved.

ACKNOWLEDGMENT

The authors acknowledge the Department of Computer Science at the University of Zululand and the Council for Scientific and Industrial Research (CSIR) for financial support.

REFERENCES

[1] N. Wollenick, "Identifying Challenges in Applying the Value Co-Creation Approach in Practice: A Case Study in the B2B Service Context," Master's thesis, Laurea University of Applied Sciences, Leppävaara, 2012.

[2] W Belmans and U. Lambrette, "The Cloud Value Chain Exposed: Key Takeaways for Network Service", Cisco, 2012.

[3] E. Luoma, "Examining Business Models of Software-as-a-Service Companies," in Economics of Grids, Clouds, Systems and Services, Springer, pp. 1-15, 2013.

[4] E. Fielt, (2010). *Business Model Canvas vs Four-Box* [Online]. Available: http://fieltnotes.blogspot.co.za/2010/09/business-model-canvas-vs-four-box.html.

[5] B. Leavy, "A system for innovating business models for breakaway growth," *Strategy and Leadership*, vol. 38, pp. 5–15, 2010.

[6] E. Fielt, "Conceptualising Business Models: Definitions, Frameworks and Classifications," *Journal of Business Models*, vol. 1, pp. 85–105, 2013.

[7] U. Shaukat, E. Ahmed, Z. Anwar and F. Xia, "Cloudlet deployment in local wireless networks: Motivation, architectures, applications, and open challenges," *Journal of Network and Computer Applications*, vol. 62, pp. 18–40, 2016.

[8] M. Satyanarayanan, P. Bahl, R. Caceres and N. Davies, "The Case for VM-Base Cloudlets in Mobile Computing," *IEEE Pervasive Computing*, vol. 8, pp. 14–23, 2009.

[9] C. R. Panigrahi, M. Tiwary, B. Pati and R. Misra, "CAB: Cloudlets as Agents of Cloud Brokers," International Conference in Advances in Computing, Communications and Informatics (ICACCI), pp.381-386, 2015.

[10] Y. Gao, W. Hu, K. Ha, B. Amos, P. Pillai and M. Satyanarayanan, "Are Cloudlets Necessary?" School of Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-CS-15-139, 2015.

[11] H. Raei and N. Yazdani, "Performability analysis of cloudlet in mobile cloud computing," *Information Sciences*, vol. *388*, pp. 99–117, 2017.

[12] M. N. Nxumalo, M.O Adigun and I. Mba, "An envisaged SLA-based Cloudlet Business Model for ensuring Service Guarantees," International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD), pp.1-4, 2018.

[13] K. Xiong and X. Chen, X, "Ensuring cloud service guarantees via service level agreement (SLA)-based resource allocation," International Conference in Distributed Computing Systems Workshops (ICDCSW), pp.35-41, 2015.

[14] E. Badidi, "A Broker-based Framework for Integrated SLA-Aware SaaS Provisioning," *International Journal on Cloud Computing: Services and Architecture*, vol. 6, pp. 01-19, 2016.

[15] D. Serrano, S. Bouchenak, Y. Kouki, T. Ledoux, J. Lejeune, J. Sopena and P. Sens, "Towards QoS-oriented SLA guarantees for online cloud services," 13th IEEE/ACM International Symposium in Cluster, Cloud and Grid Computing (CCGrid), 2013.

[16] M. C. Silva Filho, R. L. Oliveira and C. C. Monteiro, "CloudSim plus: A cloud computing simulation framework pursuing software engineering principles for improved modularity, extensibility and correctness," IFIP/IEEE Symposium in Integrated Network and Service Management (IM), 2017.

[17] N. M. P. Bocken, P. Rana and S. W. Short, "Value mapping for sustainable business thinking," *Journal of Industrial and Production Engineering*, vol. 32, pp. 88-102, 2015.

[18] R. S. Padilla, S. K. Milton and L. W. Johnson, "Components of service value in business-to-business Cloud Computing," *Journal of Cloud Computing*, vol. 4, pp. 1-20, 2015.

[19] International Integrated Reporting Council, "Value Creation background paper for <IR>," *Leadership Excellence*, pp. 1–30, 2013.

[20] M. Dara, "Value Networks and Business Models: Formulating and Demonstrating a Methodology for the Development of Value Networks and Alignment of Business Models Based on Design Science Research Methodology", MSc dissertation, University of Twente, Netherlands, 2013.

[21] M. W Johnson, A. G Lafley, "Seizing the White Space: Business Model Innovation for Growth and Renewal," Harvard Business Press, 2010.

[22] S. Gordon, "5 Ways Small Data Can Be More Valuable than Big Data, CRM Magazine," 2014. [Online]. Available: https://www.destinationcrm.com/Articles/ReadArticle.aspx?ArticleID=98927.

[23] M. Lindstrom, "The Tiny Clues That Uncover Huge Trends, Martin s Press," St. Martin's Press, 2016.

[24] Z. Pang, L. Sun, Z. Wang, E. Tian, E and S. Yang, S, "A Survey of Cloudlet Based Mobile Computing", International Conference in Cloud Computing and Big Data (CCBD), Shanghai, China, 4-6 Nov. 2015.

Development of a Wireless Intelligent Electronic Control System GSM/GPRS Unit Module Configuration for Smart Home Efficient Energy Management System

Sepeke B. Kgope
Department of Electrical and Electronic Engineering
Technology University of Johannesburg
Johannesburrg, South Africa
bernardkgope@gmail.com

Kingsley A. Ogudo
Department of Electrical and Electronic Engineering
Technology University of Johannesburg
Johannesburg, South Africa
kingsleyo@uj.ac.za

***Abstract*—The paper proposes a design and prototyping of an intelligent electronic control unit system intended for the implementation of smart home in urban and rural areas, with the intention to control and monitor energy usage by household electrical appliances. Measurement, verification of power usage, and automated efficient use of Solar power as substitute of primary power supply (Eskom power supply) in South Africa in order to reduce the high energy demand on the national grid and create low cost affordable energy to the consumers. The proposed wireless energy management system will incorporate and interface with the distribution Box (D-Box) with intelligent electronic and digital system unit with the purpose to detect presence or absence of primary power (Eskom) integrated together with solar power as alternative power supply to manage available power efficiently for free or at low cost.**

Keywords— Smart Grid, Distribution Box, Measurement and Verification (M&V), SMS, PIC, Uninterruptable Power Supply (UPS, Internet of Things (IoT)

I. INTRODUCTION

The main electric power supply utilities in South Africa, Eskom produce about 34 000 megawatts of electricity from coal to meet current demand of electricity [1], this figure is growing every year resulting to energy tariff increase. Electricity is a basic need; how will poor people survive without Eskom?

Big businesses and wealthier households may leave the Eskom grid and invest in their own electricity generation such as solar or gas [2]. Investing in alternative power generation is a good idea but people have to be at their homes to manually switch on or off their appliances while they can do the same no matter where they are.

Electricity network that can intelligently integrate the actions of all users connected to it is called a smart grid [3]. Management of energy consumption in smart grid enabled by Internet of Things (IoT) play a significant role in developing smart grid which eventually lead to low cost energy management and energy savings.

To save energy using this proposed approach, we need interconnection of embedded device to control measure and verify usage of energy wirelessly. Smart internet connectivity is becoming more interesting to internet users when intelligent electronics embedded devices are connected together to simplify energy usage systems. The household appliances control come at a cost proportional to the smart home comfort.

In this paper, we will also seek to develop a robust techniques/algorithm that will ensure software integration to the smart grid system using Internet of Things technologies. Furthermore, the advent of the development of Information Communication Technology (ICT) make life easier to most people in the cities, this paper shall present an innovative technique and low-cost energy management systems that will help bridge a gap in energy supply between people in urban and rural area alike.

II. RELATED WORK

Riaz et al. developed a home automation system which consist of a phone, a wi-fi hot spot device, a server PC and a microcontroller circuitry [9]. They designed their own circuitry board, however the introduction of a PC-server to their system increases implementation cost dramatically. Most people in the rural areas cannot afford to implement such a system.

The work of Rajesh and Bindyashree present a system which consist of a cell phone, an embedded gate (Arduino Ethernet board and Arduino Mega board) and an R transceiver (nRF2401) [10]. No server is required and therefore the costs can be kept low and affordable. They have done a good design physically with an Arduino board embedded with their software.

Basically, the transmitter transmits anything it gets on the data pins and receiver output with anything it receives irrespective of the board rate, as long as the communication is slow enough [10]. This kind of communication is very susceptible to noise and it is recommended to use a board rate of 9600b/s or less [11].

Piyare presented a system which consist of three platforms environment, a remote access device (user with android smart phone), the internet network and the home environment (router, Arduino Ethernet server, home appliances and sensors) [12]. In this kind of system, a user sends information over the internet using a cell phone application through TCP/IP, at home environment Arduino Ethernet server support TCP/IP and creates output messages in JavaScript Object Notation (JSON) format. Here, Router acts as a hotspot which means the user can also switch ON and OFF appliance at home as well remotely with its device. If the router is down the whole system can continue running using 3G or 4G as indicated in the conclusion [3]. Why

using Arduino Ethernet board and Arduino development board while own boards can be designed to meet requirements? These systems meet current technology development and implementation is expensive. Our proposed system is a standalone device using 3G and is very cheap to implement. Arduino boards will be used for modelling purpose and our final product will have our own circuitry designed boards and packaging enclosure.

III. Proposed System Layout

Fig 1. Smart Home [8]

Figure 1 above shows the proposed smart home layout with its electrical appliances. Smart Home USA website [4] defines a smart home as one that provide comfort, security, energy efficiency (low operating cost), and convenient at all time regardless whether anyone is home. This paper proposes a low-cost device which is uses Global System for Mobile (GSM) communication to switch ON and OFF, maximum of four electrical appliances remotely. In case of power failure or power restore the system intelligently detect the device shall allow user to choose which appliance/s supposed to be ON or OFF during power failure or load shedding by using Short Messaging Services (SMS) and Android cell phone application.

Some of the challenges we will seek to address through this paper work, is how we can use some of the existing electrical engineering technologies to solve the problem of scarce electricity [16].

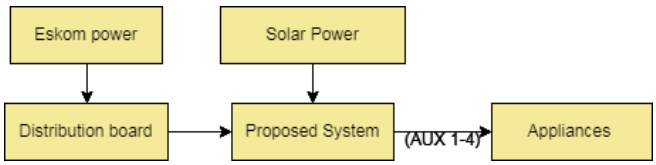


Fig 2. Block diagram for Smart home power supply

However, *figure 2* above shows that proposed system having two inputs (Eskom power and Solar power), the system will work even if there is one input preferably Eskom power. It is imperative that propose system is operated at the highest level of efficiency. In case of Eskom power failure or load shading the device shall call the user as a notification of power failure and the user shall then manage the power/energy usage from the installed solar panel or UPS. The user shall reply to choose which appliance to switch on or off from their cell phone, thereby reducing the heavy power demand on ESKOM and the customer saving substantially on the electricity bill. The approach might drastically reduce load shading if not eliminate the challenge. The user can cook supper while still at work since Eskom experience high demand between 17H00 and 22H00 during week days. The proposed shall reduce Eskom power demand during their peak hours.

The proposed system shall intelligently cut off if the load draws too much current when connected to Solar power. When there is no load shedding and power failure the system shall also notify the user if there is sufficient alternative energy and the user shall start saving by switching to available alternative energy and determining which device/appliance should be ON. The system shall also predict the point where it will automatically switch to Eskom and also notify the user.

A. Abbreviations and Acronyms

AUX – Auxiliary

ADC – Analogue to Digital Conversion

APK – Android Package

DB - Distribution Box

GSM - Global System for Mobile

MC – Microcontroller

UJ – University of Johannesburg.

PIC – Programmable Interface Controller

SMS – Short Messaging Service

B. Equations

(*Resolution of ADC) / (System Voltage) = x* / (analogue *voltage read PIC ADC pin)*

For 10 bits resolution and 5V system voltage:

1023/5V = x / (analogue voltage read)

x is decimal number to be transmitted via Bluetooth to the proposed system MC.

IV. System Implementation

This will increase the availability of skills needed by engineers, technologists, technicians as well as researchers in the following areas [16]:

- Solution oriented electrical/electronic intelligent control circuit.
- Energy measurement and verification (M&V) and efficient management.
- Alternative renewable energy (e.g.: Solar energy) as supplement to ESKOM.
- Implementation of the (IoT) technologies, in smart grid systems.

The development of network technology and communication technology make life easier to most modern people in the cities and this device shall bridge gab between metropolitan/cities and rural area technologies. It shall also

create jobs in rural areas, improve life and introduce skills and training in soldering and prototyping. Only PCBs production would be outsourced, the aim is not making dramatic money but to encourage people to follow engineering fields and give back to community.

In cities water and electricity are payable, in rural only electricity is payable. Renting is very expensive in the cities, in rural area is tenth of what is payable in city. *Table 1* below compares rural cost and city cost.

TABLE 1 : COST EFFECTIVE MEASURES

Cost	Effective Measures	
Transport	R1500	R400
Equipment	R2000	R2000
Administrative	R400	R400
Renting	Free	>R4000
Water & Electricity	R500	R900
Total	R4400	>R7700

V. METHODOLOGY & DESIGN

Devices with microcontroller are widely used in industrial field, we will design interface for connecting more than one device for parallel control. Microcontroller shall interface with GSM, Bluetooth module and 4 relays as shown in *figure 3*. The purpose of GSM is to enable the user to communicate with the device irrespective of his/her geographical area provided there is sufficient signal strength surrounding GSM (Global System for Mobile), Bluetooth module receives digital level of battery and 4 relays control appliances. In *figure 3*, GSM and SIM card send and receive SMSs through microcontroller, Bluetooth module receive ADC value from the battery read and sent by separate PIC connected to another Bluetooth module in *figure 4* below.

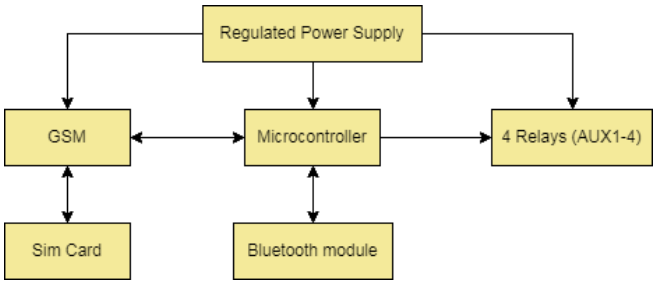


Fig 3. Block diagram of proposed system

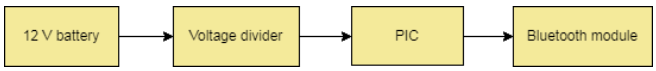


Fig 4. Battery monitoring block diagram

Bluetooth communication should only be used in occasion where there is a need for quick short-lived network communication with little concern of security [8]. Bluetooth in *figure 3 & 4* would not negatively affect the proposed system if attacked.

Fig 5. SIM900 module [13]

For the GSM module, the common SIM900 module from Sim Com shall be selected. This all-in-one module provides a standard "AT" command set for call and message management with multi-band operation for use on global GSM network carriers. As it is a commonly used GSM module, it has limited availability and is suitable for integration into custom designs.

The SIM900 is designed to be powered directly from a 3.4V-4.5V Lithium Ion battery, reducing the system cost by not requiring an additional regulator with high current. However, an unfortunate electrical characteristic of the SIM900 is a 3.1V absolute maximum voltage on its digital interface pins. As the rest of the system runs from a common 3.3V rail regulated from the battery [14]

A. Micro controllers

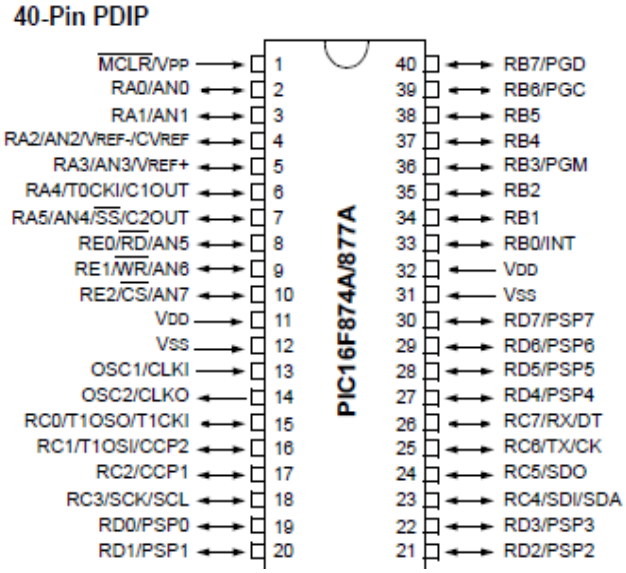


Fig 6. PIC16f877 [6]

A low-cost PIC micro MCUs have many sources of interrupt. These sources generally include one interrupt source for each peripheral module. The interrupts going to be used are [6]: Interrupt pin, USART Interrupts, Receive Interrupt, Transmit Interrupt, Data EEPROM Write Complete Interrupt and Timer1 Overflow Interrupt.

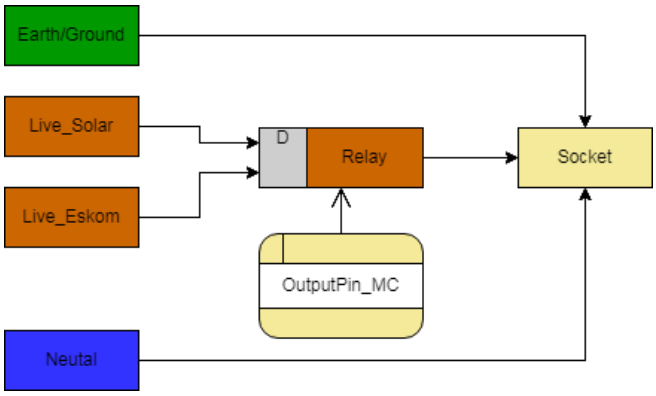


Fig 7. Relay connection

Solar power and Eskom power shall have same 'Earth' and 'Neutral', since we only switch between 'Solar live' and 'Eskom live'. *Figure 7* should be repeated 4 times to accommodate *'AUX1-4"*. Appliances would connect directly on the proposed system.

B. PROTEUS SOFTWARE (Student Vision)

This software is used to design circuit diagram, simulate the application and modelling. In addition, animate components have also been used to design the PCB layout of the project. Proteus as the ability to simulate the interaction between software running on a microcontroller and analogue or digital inputs connected to microcontroller. Virtual terminal is one of the features of Proteus Software for transmitting and receiving data outside microcontroller.

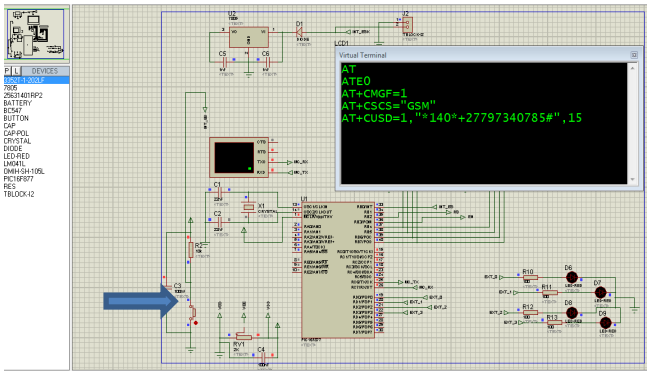


Fig 8. Eskom interrupt

Blue arrow on bottom left of *figure 8* above shows the button switch which is normally high representing the presence of Eskom. The external interrupt defined in PIC is going to raise a flag at the decreasing edge of the clock, meaning the time it goes from high to low. Regulated power supply from Eskom should be detected but external interrupt of microcontroller.

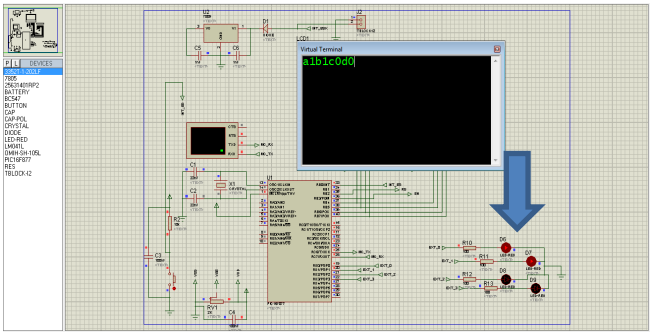


Fig 9. Switching ON/OFF devices

The blue arrow in *figure 9* shows the results obtained when manually input command of *"a1b1c0d0"* on virtual terminal and transmitted to the microcontroller. The string *"a1b1c0d0"*, means (a, b, c and d) are the devices to be switch on or off and (0/1) after letter 0 means off and 1 mean 0n. So a1b1c0d0 device 'a' and 'b' will be ON and device 'c' and 'd' will be OFF. The results show on the LEDs the output shall also be connected to the relays to indicate which power source is in use. The code test response *"Gjkja1b1C0d0kjhkj"* tested the presence of *'a1b1C0d0'* in the string and produce required results.

The SMS received is not case sensitive because each character received is converted to upper case. This keeps the project simple and understandable. Only the *table 2* logic and *table 3* characters should be learnt by the user.

TABLE 2 : LOGIC TABLE

State	Logic
ON	1
OFF	0

Each character is individually scanned, if the character is 'a' or 'A' is converted to upper case then the array of characters is created. The following character after 'A' should be number '0' or '1' if not the array is discarded and look for 'a' or 'A' again. The third character should be 'b' or 'B' if not discard the array and starts from 'a' or 'A'. The loop will continue until the array of eight characters is valid according to below table. Where 'a' or 'A' means lower case or upper case of letter 'A' and '0' or '1' means digit 0 or digit 1.

TABLE 3 : CHARACTER TABLE

	Valid characters							
Char no:	0	1	2	3	4	5	6	7
Char	a/A	0/1	b/B	0/1	c/C	0/1	d/D	0/1

C. Android Studio developer[15]

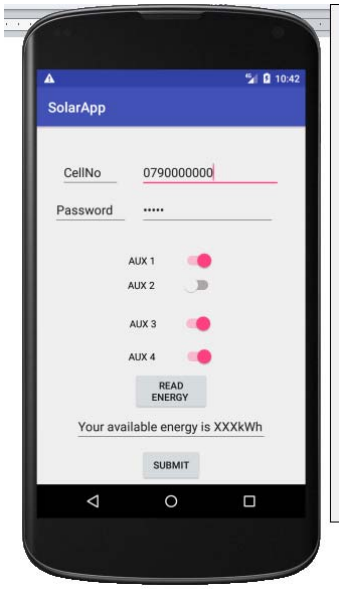


Fig 10. Android Solar App

The user shall enter cell number and password, if password is correct the Solar App will send an SMS to the GSM connected to the SIM card with cell number entered. *'Read Energy button'* in *figure 10* above will read available power on the batteries display the value above *'Submit button'*. *AUX (1 - 4)* represent appliance connected on the system similar to device *"a, b, c, and d"* in *Table 3* and shall only be active after *'submit button'* is pressed.

An ordinary phone can also send an SMS to the device and achieve the same results without reading available energy. Android Studio [15] offers even more features that enhance your productivity when building Android apps, such as: a fast and feature-rich emulator, a unified environment where you can develop for all Android devices, instant run to push changes to your running app without building a new APK, code templates and GitHub integration to help you build common app features and import sample code, testing tools and frameworks, Lint tools to catch performance, usability, version compatibility, and other problems, support C++ and a built-in support for Google Cloud Platform, making it easy to integrate Google Cloud Messaging and App Engine.

VI. Conclussion

The proposed intelligent electronic control system uses both internal and external interrupts of microcontroller (MC). External interrupt handles the presence and absence of Eskom power which is binary 0 or 1, internal interrupts handles the presence of 8 bits transmitted or received character in the buffer memory of MC. Internal timer interrupt of MC connected to battery which charge through Solar power, use 16 bit timer to count 500uS hundred times to make 50mS and read analogue value external battery voltage and covert it digital using *'B. Equation'* under *'III Propose System'* and transmit it to the proposed system connected to Solar panel and inverter. The receiver interrupt has higher priority than other interrupts in this proposed system.

The availability of free services which most of our service providers provide will make the proposed device to be cheaper. For an example Vodacom give one chance to personalize *'Please Call Me'* with maximum name of 10 characters per day and also allow the subscriber to send about ten *'Please Call Me'* every day. This means the user will have one free chance to control electrical appliance per day. In conclusion, this paper proposes a cost-effective device and mass production will make it affordable to everyone. In the future, we intend to develop a PHP web server with MySQL database to monitor and verify power usage of electrical appliance and calculate the amount saved when using solar power source.

References

[1] Eskom. "Generating Electricity," Mar. 15, 2018. [online] Available: http://www.eskom.co.za/AboutElectricity/ElectricityTechnologies/Pages/Generating_Electricity.aspx. [Accessed: Mar. 15, 2018]

[2] Mathew le Codeur, "Eskom accussed of hamming on economy by selfishly focusing on itself" Oct 30, 2017, [Online] Available: https://www.fin24.com/Economy/Eskom/eskom-accused-of-harming-economy-by-selfishly-focusing-on-itself-20171030. [Accessed: Mar 18, 2018]

[3] Bugess, M. & Rmanov, A. "What is the internet of things?". [Online] Available: http://www.wired.co.uk/article/internet-of-things-what-is-explained-iot. [Accessed Mar 20, 2018]

[4] Smart Home USA, "What is Smart Home", Sep 2, 2018. [online] Available: https://www.smarthomeusa.com/smarthome. [Accessed Sep 2, 2018]

[5] Business Connection (pty) Ltd, "Cloud Computing", 2017. [Online] Available: http://www.wired.co.uk/article/internet-of-things-what-is-explained-iot [Accessed Mar 15, 2018]

[6] Microchip, "PIC16F87XA Datasheet" [online] Available http://ww1.microchip.com/downloads/en/DeviceDoc/39582b.pdf [Accessed on Mar 17, 2018]

[7] R. Gupta and M. Awad, "PIC Controller Based Process Control System using GSM", *International Journal of Advanced Research in Computer Engineering & Technology*, vol. 1(5), pp. 2278-1323, 2012.

[8] V. S. Gunge and P. S. Yalagi, "Smart Home Automation: A Literature Review", *International Journal of Computer Applications*, 2016.

[9] M. T. Riaz, E. M. Ahmed, F. Durrani and M. A. Mond, "Wireless Android Based Home Automation System", *Advances in Science, Technology and Engineering Systems Journal*, vol. 2(1), pp. 234-239, 2017.

[10] K. R. Rajesh and C. A. Bindyashree, "Multiple Appliances Controlling and Monitoring System based on Wireless Embedded Home Gateway", *International Journal of Innovative Research in Computer and Communication Engineering*, vol 3(4), pp. 2872-2877, 2015.

[11] C. Amariei, Arduino Development Cookbook. Birmingham: Packt Publishing, 2015.

[12] R. Piyare, "Internet of Things: Ubiquitous Home Control and Monitoring System using Android based Smart Phone", *International Journal of Internet of Things*, vol. 2(1), pp. 5-11, 2013.

[13] SIM900 The GSM/GPRS modulefor M2M application [online] Available: http://www.propox.com/download/docs/SIM900.pdf [Accessed Mar 17, 2018].

[14] AT03821, SAM D20 GPS Tracker Reference Design, Atmel, San Jose, 2014.

[15] Meet Andriod studio, [Online] , Available: https://developer.android.com/studio/intro/index.html, [accessed Mar 15, 2018].

[16] Ouahada K,. Smart Home presentation, University of Johannesburg, Mar 17, 2018.

Artificial Intelligence in Cyber Threats Intelligence

Roumen Trifonov
Faculty Computer Systems and Technologies
Technical University of Sofia
Sofia, Bulgaria
r_trifonov@tu-sofia.bg

Ognyan Nakov
Faculty Computer Systems and Technologies
Technical University of Sofia
Sofia, Bulgaria
nakov@tu-sofia.bg

Valeri Mladenov
Faculty Automation
Technical University of Sofia
Sofia, Bulgaria
valerim@tu-sofia.bg

***Abstract*—In the field of Cyber Security there has been a transition from the stage of Cyber Criminality to the stage of Cyber War over the last few years. According to the new challenges, the expert community has two main approaches: to adopt the philosophy and methods of Military Intelligence, and to use Artificial Intelligence methods for counteraction of Cyber Attacks. This paper describes some of the results obtained at Technical University of Sofia in the implementation of project related to the application of intelligent methods for increasing the security in computer networks. The analysis of the feasibility of various Artificial Intelligence methods has shown that a method that is equally effective for all stages of the Cyber Intelligence cannot be identified. While for Tactical Cyber Threats Intelligence has been selected and experimented a Multi-Agent System, the Recurrent Neural Networks are offered for the needs of Operational Cyber Threats Intelligence.**

Keywords— Cyber Threats Intelligence, Artificial Intelligence, Behaviour Assessment, Neural Networks, Sequential Feature Selection, Remote Network Monitoring

I. INTRODUCTION

Over the last few years, the trends of transition of the Cyber Threats from the Cyber-Crime phase to the Cyber-War phase has also prompted an adequate transition of Cyber Defense techniques to military technology [1]. First of all, this concerns the perception in the analysis of the threats of the so-called "Cyber Kill Chain" model, as well as the application of traditional Military Intelligence Technology.

Furthermore, in the conditions where well-resourced and trained adversaries conduct multi-year intrusion campaigns targeting highly sensitive economic, proprietary, or national security information, the network defense techniques which leverage knowledge about these adversaries can create an intelligence feedback loop, enabling defenders to establish a state of information superiority which decreases the adversary's likelihood of success with each subsequent intrusion attempt. That's why according to the vast majority of experts, the qualitative transition to new Cyber Defense tools must involve the widespread use of Artificial Intelligence methods to analyze information exchanged, network flows, sources of threats, and to plan effective impact measures, including proactive ones.

Following these trends, the Faculty of Computer Systems and Technology at Technical University of Sofia began research on the application of intelligent methods for increasing the security in computer networks. An essential section of this investigation is dedicated to the Cyber Threat Intelligence. The present article summarizes some results of a research done by the project team.

II. BASIC FEATURES OF THE CYBER THREATS INTELLIGENCE PROBLEM FORMULATION

The Cyber Intelligence or, more precisely, Cyber Threats Intelligence (CTI) has the following definition in the draft Bulgarian National Cyber Security Strategy [2]:

- establishment of mechanisms and technical means to maintain an up-to-date picture of possible threats of different scale, sources and character, trends in geopolitical context development and relevant national cyber picture analysis and;
- development of capabilities to help identify attribution sources and take appropriate forms of protection and counteraction.

According to the documents of INSA (Intelligence and National Security Alliance) [3, 4, 5] the preparation of the intelligence in cyber operational environment is a systematic and continuous process of analyzing potential threats to detect a suspicious set of activities that may endanger systems, networks, information, employees, or customers by providing means to visualize and evaluate a number of specific penetration sensor inputs to bring up a particular threat. This process supports the organization's risk management strategy and decision-making in the area of information security. Its application identifies potential threats and assists security and risk managers selectively implement and maximize deep defense strategies by better understanding the critical points in time and space in the operating environment.

The Cyber Threats Intelligence Cycle [6] is a systematic, continuous process of analyzing potential threats to detect a suspicious set of activities that might threaten the organization's systems, networks, information, employees, or customers by providing a means of visualizing and assessing a number of specific intrusion sensor inputs and open source information to infer specific threat courses of action. The model supports the organization's risk management strategy and the information security group's decision-making. The application of the model identifies potential threat courses of action and helps the security and risk management leaders selectively apply and maximize a defence in depth strategy via a greater understanding of the organization's cyber threats at critical points in time and space in the operational environment by:

a) defining the operational environment;

b) describing the operational environment effects on network defense;

c) evaluating the cyber threats, and

d) developing cyber threat courses of action

Fig. 1 is a graphical representation of the Cyber Threat Intelligence Cycle.

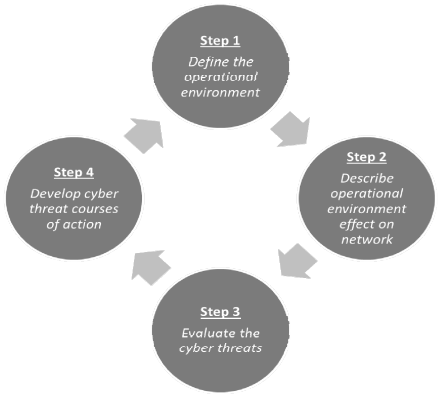


Fig. 1. Cyber Threat Intelligence Cycle

Like its military analogue, the Cyber Threats Intelligence is developed at three levels: strategic, operational, and tactical. For the purposes of this study, the second one is considered: INSA defines [5] the operational level as: "The level at which campaigns and major operations are planned, conducted, and sustained to achieve strategic objectives within theaters or other operational areas. At this level, actors build the capabilities needed to support the tactical operations. They maneuver in cyberspace to position capability where they need to in order to be effective in their tactical missions. At the operational level, an organization's operating environment can be described in terms of physical, logical, information and social layers".

III. METHODS OF ARTIFICIAL INTELLIGENCE IN NETWORK AND INFORMATION SECURITY

The essence of Artificial Intelligence (AI) is based on the statement that people's intelligence (the potential (inborn) ability of a conscious individual to conclude on a given information) can be described so precisely that it is machine-simulated. After several decades of research, AI is not only the subject of research or planning of some movement, but also of more complex and interdependent solutions. Artificial Intelligence is defined as the intelligence displayed by machines and / or software. This is an academic field of study exploring the goal of creating intelligence. The main issues explored by AI include reasoning, knowledge presentation, automated planning and scheduling, machine learning, natural language processing, computer vision, robotics and common intelligence.

AI enables us to develop autonomous computer solutions that adapt to their context of use, using self-management, self-tuning and self-configuration, self-diagnosis and self-healing. When it comes to the future of information security, AI looks like a very promising field of research that focuses on improving cyberspace security measures.

With rapid pace of development and the desire for more effective countermeasures, Artificial Intelligence comes as a natural solution to the problem of coping with the ever-growing number of network attacks. Applications in the field of AI are widely accepted by the modern information society. This interdisciplinary endeavor has created a joint link between computer specialists and network engineers in designing, simulating and developing network penetration patterns and their characteristics.

As mentioned in the introduction to this article, world practice has already noted a significant number of various Artificial Intelligence applications in computer security. Without trying for a comprehensive classification, we could divide these methods into two main directions:

A. Conditionally named "distributed" methods:

A1. Multi-Agent Systems of Intelligent Agents;

A2. Neural Networks;

A3. Artificial Immune Systems and Genetic Algorithms.

B. Conveniently named "compact" methods:

B1. Machine Learning Systems, including: associative methods, inductive logic programming, Bayes classification;

B2. Pattern recognition algorithms;

B3. Expert Systems;

B4. Fuzzy logic.

Having into account this variety of methods, it is of particular importance that adequate criteria are selected for the assessment and selection of a specific application for each specific solution. In the above mentioned project, the specification was carried out for two of the main sections of CTI. It should be noted here that within the project the application of Multi-Agent systems was chosen and experimented as the most appropriate method for the needs of the Tactical Cyber Intelligence.

IV. METHODS OF ARTIFICIAL INTELLIGENCE SUITABLE FOR OPERATIONAL CYBER THREATS INTELLIGENCE

The ultimate goal of Operational Cyber Intelligence is to reduce risk to an organization's critical mission and assets by: defining the operating environment; describing the impact of the operating environment; evaluating the adversary; and determining potential adversarial courses of action (COA). The Operational Cyber Intelligence provides a thread that links the probability and impact of a cyber attack with its strategic level implications by ensuring a coherent framework for analysis and prioritization of potential threats and vulnerabilities given the organization's threat environment. Operational Intelligence is based on the Doctrine of Active Defense. Instead of searching for information regarding a specific attack against the organization, it focuses on analyzing the opponents' combat doctrines, weapon systems and attack and operational scenarios. This approach shifts the center of gravity to the ability to respond and block the outcome of the attack within the organizational environment or in its immediate vicinity.

Our main idea was that the basis for the automation of the Operational CTI can be the behavioral model of the likely adversary. It should be emphasized that the problem of using artificial intelligence methods in the Operational CTI is a completely new matter, and systematized literary sources have not yet been found. Only, there are reports concerning the use of behavioral analysis based on machine learning by

the companies: Exabeam (USA), Darktrace (UK), CyberX (USA), Interset (Canada).

The TU-Sofia team concluded that the activity and the outgoing traffic in the network of the supposed adversary were to be the main source of information for building his behavioural model. This evokes analogies with the Non-Invasive Brain - Computer Interface whereby the physiological signals of the human brain (for example, through Electroencephalograms (EEGs)) can be used for human emotions evaluation [7].

Indeed, the streams of measured parameters received by n-number different IP addresses of the monitored object using RFC 1757 Remote Network Monitoring methods [8] can be compared to EEG with n-number of channels.

If this analogy is applied in practice, first of all, on the order of the classification model of emotions [9], a basic classification of the behavior of the possible adversary, based on the needs of our research, must be constructed,. Currently, in the absence of references for such studies, it is assumed that this behavior can be divided for the present into two basic types: hostile and non-hostile.

In order to obtain the best possible performances, it is necessary to work with a smaller number of variables which describe some relevant properties of the data retrieved from the network. These variables are known as "features". Features can be aggregated into a vector known as "feature vector". Thus, feature extraction can be defined as an operation which transforms one or several signals into a feature vector. Identifying and extracting good features from signals is a crucial step, because otherwise the classification algorithm will have trouble identifying the class of these features, i.e., the behavioral state of the possible adversary. According to some researchers [10], it seems that the choice of a proper pre-processing and feature extraction method have more impact on the final performances than the selection of a good classification algorithm [16].

Therefore, following the analogy of the Brain-Computer Interface, two basic tasks have to be solved:

- to find a suitable approach to selecting characteristics from which to derive features suitable for behavioral interpretation and validation. In doing so, the necessary inter-subject discrimination of the features for the subsequent classification must be ensured;
- to build and optimize an ensemble of classifiers based on trained models to be used to assess behavior.

According to the researcher's scenario, design of the system of assessing the behaviour of the supposed adversary can consist of two main phases: 1) offline training phase to calibrate the system and 2) online phase which uses the system to recognize the type of behavior states and translate them into the computer commands. Both offline and online phases follow a closed-loop process, generally composed of six steps:

a) network activity measurement- this step consists in network surveillance of broadband Internet traffic (e-Mails, Web traffic, instant messengers, etc.) using methods, such Packet Capture Appliances Fig. 2 in order to obtain signals reflecting the opponent's intentions [11];

b) preprocessing - this step consists in cleaning and denoising input data to enhance the relevant information embedded in the signals;

c) feature extraction – this extraction aims at describing the signals by a few relevant values called "features";

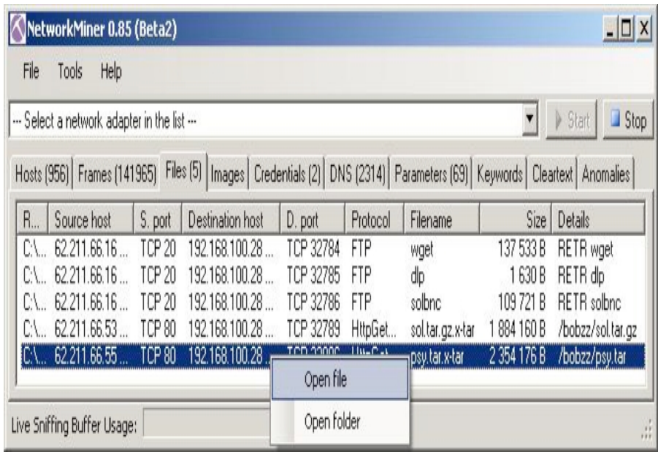


Fig. 2. Packet Capture Appliances

d) classification - this step assigns a class to a set of features extracted from the signals, which corresponds to the kind of behavioral state identified. This step can also be denoted as "feature translation". Classification algorithms are known as "classifiers" [15];

e) translation into a command/application - once the behavioral state is identified, a command is associated with this state in order to control a given application.

Once the data have been acquired, they are pre-processed to clean (de-noise) the signals and to enhance relevant information embedded in these signals. The pre-processing step aims at increasing the signal-to-noise ratio of the input signals.

To perform this pre-processing, various spatial-spectro-temporal filters [10] can be used. Naturally, numerous other pre-processing methods, which are more complex and more advanced, can be proposed and used. But in our initial experiments we were based on two of the most popular methods, namely, Independent Component Analysis (ICA) and Common Spatial Patterns (CSP) method.

Based on a study of literary sources, the Echo State Network (ESN) method was proposed as a mechanism for feature selection – this is a class of Recurrent Neural Networks where the so-called "Reservoir Computing" approach for training is formulated [12].

The basic structure of an ESN, presented in Fig. 3, consists of a reservoir of random connected dynamic neurons with sigmoid nonlinearities (usually hyperbolic tangent):

$$r(k) = f\,res(W\,inin(k) + W\,resr(k-1))$$

and a linear readout *f out* (usually identity function) at the output:

$$out\,(k) = f\,out\,(W\,out\,[in\,(k)\,r\,(k)])$$

Here *k* denotes discrete time instant; *in(k)* is a vector of network inputs, *r(k)* - a vector of the reservoir neurons states and *out(k)* – a vector of network outputs; *nin*, *nout* and *nr* are the dimensions of the corresponding vectors *in*, *out*

and *r* respectively; *W out* is a trainable *nout* (*nin*+*nr*) matrix; *W in* and *W res* are *nr nin* and *nr nr* matrices that are randomly generated and are not trainable. In some applications, direct connection from the input to the readout is omitted.

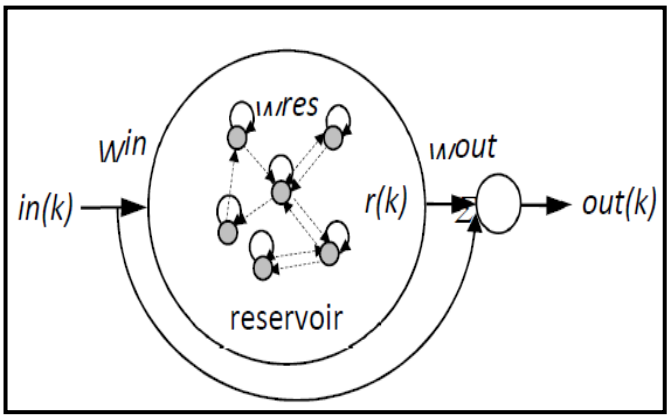


Fig. 3. Basic structure of an ESN

The main advantage of the ESN is the simplified training algorithm since only weights of the connections from the reservoir to the readout neurons are subject to training [13]. Thus instead of gradient descent learning much faster least squares method can be used.

We started on the presumption that using reservoir computing pre-training is beneficial for selecting the most relevant discriminative features and reaching state-of-the-art performance for subject independent recognition. The Reservoir Computing approach could be used not only for time series processing but also for high dimensional static data representation. Finally, the existing practice shows that IP-trained ESNs outperform pre-trained deep auto-encoders and can actually achieve almost 100% testing accuracy.

Exploring the feasibility of training cross-subject classifiers, we have settled on the Sequential Feature Selection (SFS) procedure [14] that reduces the inherent data variability and can lead to a high inter-subject behaviour status recognition accuracy. Starting from an empty set, SFS increments sequentially a new feature that best predicts the class at the current iteration. The process stops when there is no more improvement in the prediction. SFS is a very effective way to identify the dominant behavioral signatures across subjects. However, it is a computational heavy and time-consuming procedure, which was the main motivation to look for a computationally less intensive alternative.

The state of the art of the works described in this article can be defined as a transition from the development of a theoretical model to an experimental setting.

As the experiments are in their early stage, it is necessary to point out that the results are encouraging, but it is still too early to declare any definitive conclusions.

V. CONCLUSION

As can be seen from the above, the process of introducing Artificial Intelligence methods at the different levels of Cyber Threat Intelligence is at very different stages: while in Tactical Intelligence, it has long gone out of the phase of research and experiments and is used for building real effective systems, In the field of Operational Intelligence, these studies are in a very initial phase and require the commitment of substantial resources. Furthermore, the question arises as to the application of possible outcomes of Operational Iintelligence in the activity of Tactical Intelligence systems, which are intended to neutralize the immediate threats to computer systems and networks.

REFERENCES

[1] *ENISA Threats Landscape Report* 2016: 15 Top Cyber-Threats and Trends, ENISA, 2017

[2] Republic of Bulgaria: *National Cyber Security Strategy "Cyber Resilient Bulgaria 2020"*, 2016-03 NCSS Bulgaria final draft v 5 3, Bulgarian government, 2016

[3] *Cyber Intelligence: Setting the Landscape for an emerging Discipline, Intelligence and National Security Alliance* , INSA, 2011

[4] *Operational Level of Cyber Intelligence*, INSA, 2013

[5] *Operational Cyber Intelligence*, INSA, 2014

[6] Brian P. Kime *Threat Intelligence: Planning and Direction*, SANS Institute, 2015

[7] Liu Y., Sourina O. and Nguyen M. K., Real-time EEG-based human emotion recognition and visualization, in *Proceedings of the International Conference on Cyberworlds* (CW '10), 2010, Singapore

[8] [RFC 1757] Remote Network Monitoring Management Information Base, Carnegie Mellon University, 1995

[9] L. Bozhkov, P. Georgieva Classification models of emotional biosignals evoked while viewing affective pictures, in *Proceedings of the International Conference on Simulation and Modeling Methodologies, Technologies and Applications* (SIMULTECH), 2014, Vienna

[10] Hammon P.S. and Sa V.R. de, Preprocessing and meta-classification for brain-computer interfaces, *IEEE Transactions on Biomedical Engineering*, 54(3), 2007.

[11] Erik Hjelmvik, Passive Network Security Analysis with Network Miner, (IN)SECURE Magazine, no. 18, pp. 18–21, 2008.

[12] Lukosevicius M. and Jaeger H., Reservoir computing approaches to recurrent neural network training, *Computer Science Review*, vol. 3, 2009

[13] Koprinkova-Hristova P., Bozhkov L. and Georgieva P., Echo State Networks for feature selection in affective computing, in *Proceedings of the 13th International Conference on Practical Applications of Agents and Multi-Agent Systems (PAAMS)*, 3-5 June 2015, Spain

[14] Guyon I. and Elisseeff A. An Introduction to Variable and Feature Selection, *Journal of Machine Learning Research*, vol. 3, 2003.

[15] Hanan Hindy, David Brosset, Ethan Bayne, Amar Seeam, Christos Tachtatzis, Robert Atkinson, and XavierBellekens. 2018. A Taxonomy and Survey of Intrusion Detection System Design Techniques, Network Threatsand Datasets. 1, 1 (June 2018), 35 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

[16] Elike Hodo, Xavier Bellekens, Ephraim Iorkyase, Andrew Hamilton, Christos Tachtatzis and Robert Atkinson. Machine Learning Approach for Detection of nonTor Traffic. Journal of Cyber Security, Vol. 6.2, pp 171–194, November 2017. doi: 10.13052/jcsm2245-1439.624

Parallel Image Stitching Based on Multithreaded Processing on GPU

Ognyan Nakov
Faculty of Computer Systems and Technologies
Technical University
Sofia, Bulgaria
nakov@tu-sofia.bg

Elizabet Mihaylova
Faculty of Computer Systems and Technologies
Technical University
Sofia, Bulgaria
elizabeth.mihailova@gmail.com

Milena Lazarova
Faculty of Computer Systems and Technologies
Technical University
Sofia, Bulgaria
milaz@tu-sofia.bg

Valeri Mladenov
Faculty of Theoretical Electrical Engineering
Technical University
Sofia, Bulgaria
valerim@tu-sofia.bg

***Abstract*—The paper discusses multithreaded processing of images on graphic processing units for the purposes of feature detection and matching. The problem of feature detection and feature correspondence is applied for image stitching and panorama creation. Parallel GPU implementation based on nVidia CUDA is presented and experimentally evaluated and compared by parallel multithread CPU processing for shared memory parallel computational model.**

Keywords—feature detection, feature matching, general purpose computations on GPU, multithreading,image stitching

I. INTRODUCTION

In computer vision and image processing feature detection is an important operation applied on single image as well as multiple images as part of the processing stages for image analyses and information extraction [1]. Feature could be any piece of information that helps solving a computational task. For example, feature could be point, edge or object in an image or another specific structure. Feature also is the result form feature detection. Based on the detected features numerous different tasks for understanding the image content can be solved. Among these are depth analyses and 3D model reconstruction, image stitching and panorama image creation [2, 3], augmented reality[4, 5]

Image stitching is aimed at the alignment of a set of multiple images in order to generate a composite panorama image. Analyses of the content of multiple images to detect and match common features of the contained objects is usually applied for solving image stitching problem [6]. Most of the image stitching approaches involve pictures that well overlap and are taken with similar exposures [7, 8], although in cases such as by doing high dynamic range imaging different exposures may benefit [9, 10, 11] Similar sequence of steps is also applied for object detection and 3D reconstruction using multi-view geometry analyses of set of images [12, 13]. Image stitching is often used for object insertion, super resolution [14] and texture synthesis [15], image matching and recognition [16]

Graphics processing unit is a special purpose device aimed at fast processing of geometry models and image data for generation and visualization of graphical information. The hardware architecture of the recent GPUs accommodates highly parallel capabilities of stream processing. Provided the possibility to reprogram and alter the predefined functionality of the GPU by running user specified function on the graphics devices turns them into high-performance computational systems. Utilization of graphics processing units for general purpose computations (GPGPU) lead to a steady trend in top supercomputer systems: the most powerful machines worldwide are hybrid high-performance systems utilizing both CPU and GPU processing power [17].

The paper is aimed at presenting a multithreaded processing for feature detection and feature matching in multiple images. The utilization of parallel processing using multiple threads running on the GPU is presented and discussed. Parallel GPU implementation based on nVidia CUDA for the purposes of image stitching and panorama image creation is presented and experimentally evaluated and compared by parallel multithread CPU processing for shared memory parallel computational model.

II. NVIDIA CUDA ARCHITECTURE

CUDA (Compute Unified Device Architecture) is a parallel hardware and software architectural model and framework that allows utilization of the computational power of GPU by running programs written in C, C++, OpenCL, DirectCompute [18].

CUDA extends the single thread model by minimal abstractions in order to deliver parallel multithreaded computational models aimed at different problems and tasks.

The abstract parallel model comprises both task-level parallelism and coarse data-level parallelism. The later can be further regarded as thread level parallelism and fine granular datalevel parallelism.

In order to maximize the efficiency of the parallel computational model that will run on the GPU utilizing CUDA it is necessary first to divide the task into independent

subtasks that are mapped to a grid of blocks and second, to divide each of the tasks in turn into subtasks that can be executed in parallel by the threads in each block. Provided the problem solved can be regarded as composed of an adequate number of independent subtasks, the CUDA model allows for better performance and increased productivity as independent subtasks can be executed simultaneously on the available hardware device thus providing automatic scalability (Fig. 1).

CUDA is aimed at running heterogeneous computational models with serial code executed on the host computer utilizing the CPU and the parallel code running as parallel kernel function on the graphics device utilizing the thread level parallelism of each stream multiprocessor of the GPU (fig.2). GPU threads are lightweight, their lifecycle and switching time are minimized so that thousands of threads can be created and controlled for few CPU cycle.

Since the host computer and the graphics device are regarded as separate devices by CUDA, it allows for simultaneous execution of the serial and the parallel part of the code without requirement for memory management and data races prevention. A set of blocks that execute one and the same kernel function and have access to data in global memory is regarded as grid.

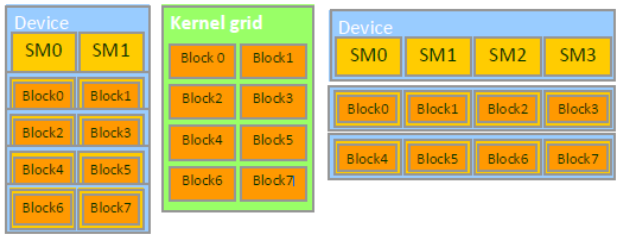


Fig. 1. Automatic scalability [19]

Several parameters of the block are programmatically specified:

- block size – from 1 to 512
- concurrent threads;
- block dimensions – 1D, 2D, 3D.

All threads in one block run the same kernel function. Threads have unique identifier in the block. Threads in one and the same block can share data in the global device memory and can be synchronized. Threads from different blocks in one and the same grid can cooperate through atomic operations in the global memory shared by the grid threads.

The number of blocks in each grid and their order is set on kernel function call: gridDimx and gridDimy determine the number of blocks in x and y dimension respectively and can vary between 1 and 65 536. The number of blocks cannot be dynamically modified.

Thus each thread uses its identifier in the block as well as block identifiers in the grid in order to specify the tasks and the data partition that the thread will work on.

So, when a kernel function runs, a set of grids and threads is started on the graphics device (Fig.3). The number of blocks of threads that are executed on the device depends on the device parameters and thus provides transparent scalability of the multithreaded computations. For better utilization of the device hardware the number of blocks should equals the number of the multiprocessors. Each multiprocessor consists of stream processors and uses SIMT (Single Instruction Multiple Threads) mode of processing: each thread is executed concurrent to the other threads and has own register file, call stack and local memory [20].

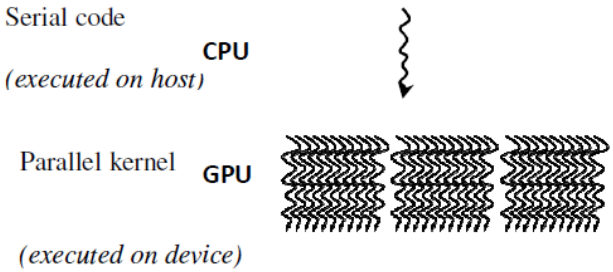


Fig. 2. Heterogeneous computational model

CUDA architecture supports several levels of parallelism:

- thread parallelism: each thread is independent;
- task parallelism: different blocks of threads are independent; independent kernels;
- data parallelism: across threads in a thread block; across thread blocks in a kernel.

This provides possibilities for effective utilization of the possibilities for speeding up the computations of particular parallel algorithm exploiting both levels of parallelism

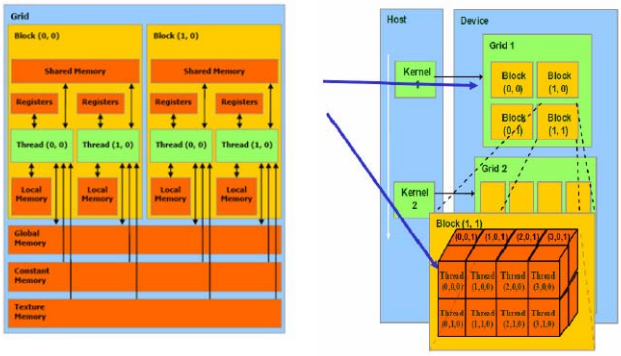


Fig. 3. nVidia CUDA architecture [19]

III. IMAGE STITCHING

Image stitching methods are broadly classified into two main methods: direct, and feature based. Direct method for image stitching takes and analyse every single bit of infomation in the image. Although it is the more accurate method of the two, because it uses all of the available data, direct method is considered slower because of the reasons that most of the times it requires a human operator to interact with the system and to define the

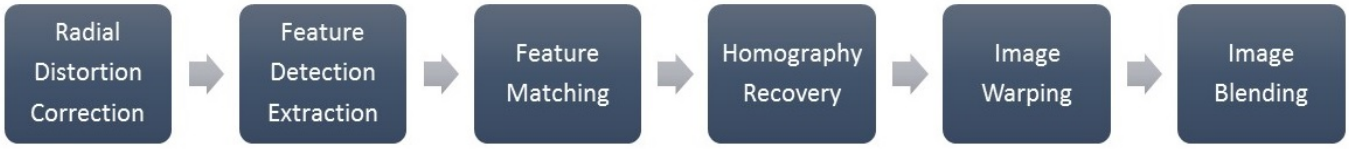


Fig. 4. Image stitching pipeline

matching image parts on the one hand, and on the other it takes more time for computation. The feature-based methods do not use all data in the image but instead focused on specific structures in the image such as ponit, edge, object and assume that the other regions in the image are consistent and does not involve a human input.

The process of creating panoramic images based on a set of images (also known as image stitching) comprises a sequence of steps that can be divided in two stages [21]: registration of images and image composition. The image stitching pipeline is presented in Fig. 4.

The registration stage requires feature detection in the set of multiple images and matching the detected features. Techniques for implementing the images registration can be regarded as: (1) intensity based in which pixels correspondence is evaluated based on the pixel's intensity values and (2) feature based which are based on extraction and matching of scale and space invariant features. Features that can be used for the image registration include Harris and Moravec corner detectors, Canny edge detector, Hough transform. Among the most widely used approaches for registration in image stitching pipeline are SIFT (Scaleinvariant feature transform) [22] and SURF (Speeded Up Robust Feature) [23]. The feature matching step of the image registration requires the feature descriptors to be match for best correspondence evaluation. Based on the matching stage image homography is recovered and used for panorama composition. The homography recovery is aimed at estimating the camera parameters and can utilize Least Median of Squares (LMS) of k-neighbors search algorithms or RANSAC (RANdom Sample Consensus) algorithm [24]. At the stage of image composition a projective transformation for image adjustment is used with additional alignment for gap closing and parallax removal.[25, 26] It is preferred to use technics for paralax removal when 3D reconstruction is not applied and/or dense sampling of a scene. If upmentioned techniques are applied there are multi-perspective panorama techniques that could handle parallax well.[27, 28, 29, 30] Final image is composed with image blending using Gaussian smoothing, Laplacian pyramids and Gradient domain blending. The last one is widely used in compression of dynamic range, separation of images to layers, image editing, image inpainting [31, 31, 33, 34, 35, 36, 37, 38]

The image stitching processing pipeline can be summarized as comprising the following steps:

1) Feature detection using SURF*;*

2) Homography estimation by feature descriptors matching using nearest neighbor search*;*

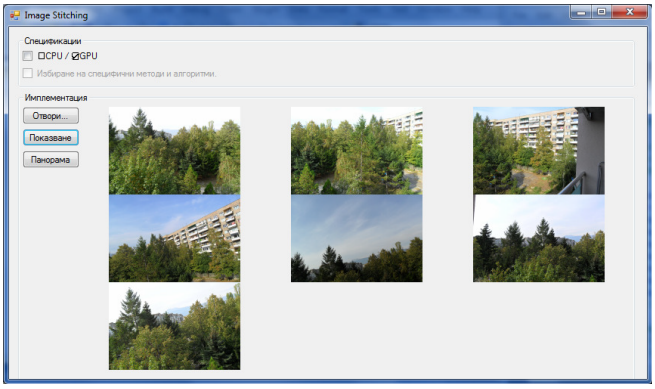
3) Image warping and deformation correction;

4) Panorama composition by image blending, color correction and stitches removal.

IV. Experimental results

Image stitching processing stages were implemented as an application capable of both processing using multithreaded CPU or using multithreaded GPU computations. The implementations utilizes OpenCV library modules [39,40]. The CPU implementation utilizes multithreading based on OpenMP API [41]. The user interface of the application provides for choice of CPU or

GPU execution as well as algorithm and parameter selection for the stitching pipeline stages, (Fig. 5).



(a)

(b)

Fig. 5. Image User interface (a) and example result (b) of pipeline

The results of the speedup of the GPU execution over CPU achieved for several sets of images (from 3 to 10 images per set) of resolution 3984 x 2656 are given in Table 1. Comparison of the time required for each processing stage of the image stitching pipeline between GPU and CPU implementations is shown on Fig. 6 and the speedup of the GPU processing at each stage is presented on Fig.7. The experimental results show that the overall speedup achieved is 7,5 times faster execution of the parallel implementation using multithreaded processing on the GPU. For some of the processing stages – feature matching and deformation correction, GPU execution is 9 to 10 times faster than the CPU run of the relevant stage due to the utilization of the data parallelism of the relevant algorithms when mapped to the thread hierarchy of the nVidia CUDA architecture. Feature detection and panorama image blending stages also show better performance when run on the GPU– speedup is

respectively 2 and 5 for these processing steps. For the image warping and color correction the graphics device implementation give similar or worse results for the time spend on these stages.

TABLE I. SPEEDUP OF THE PARALLEL GPU PROCESSING

Processing stage	Speedup
Feature detection	2,13
Feature matching	10,41
Rotation	1,24
Geometric distortion	1,04
Deformation correction	9,13
Color correction	0,95
Stitches detection	1,14
Final panorama composition	4,92
Total image stitching	7,49

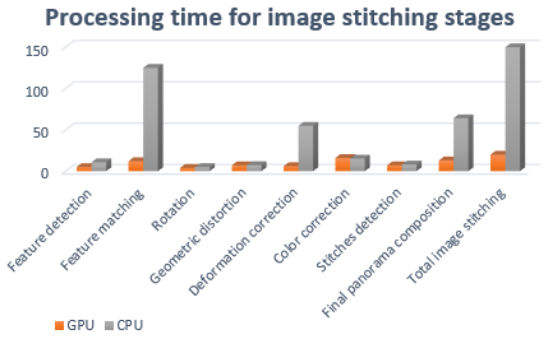


Fig. 6. Processing time for the image stiching on CPU and GPU

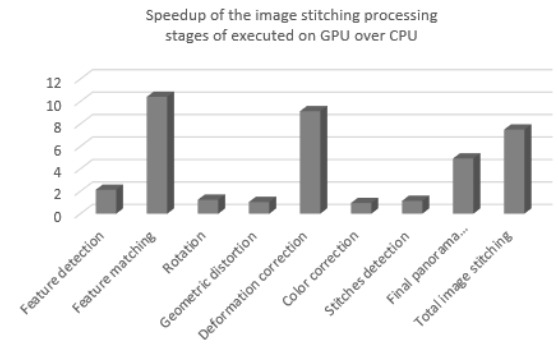


Fig. 7. Speedup of the image stitching processing stages of executed on GPU over CPU

V. CONCLUSION

The paper presented a multithreaded processing for image stitching and panorama image creation using multiple images. The possibilities for utilization of data and task parallelism using multiple threads running on the GPU is presented and discussed. The developed implementation of image stitching utilizes parallel processing on the GPU based on nVidia CUDA and on the CPU based on OpenMP API. The experimental results show that the GPU processing outperforms the CPU execution by more than 7 times better performance parameters. The speedup is up to 10 for the stages of feature matching and deformation correction that best utilizes the thread mapping of the parallel algorithm to the graphics device architecture. Further improvement of the results can be achieved by speeding up the time consuming processing stages using a hybrid CPU-GPU parallel model with simultaneous computations both on the device and on the host computer.

REFERENCES

[1] E. Davies, Machine Vision, Morgan Kaufmann, 2005

[2] C. Wöhler, 3D Computer Vision: Efficient Methods and Applications, Springer, 2009.

[3] B. Cyganek, J. Siebert, An introduction to 3D computer vision techniques and algorithms, Wiley, 2009.

[4] Gabriel Takacs, Vijay Chandrasekhar, Natasha Gelfand, Yingen Xiong, Wei-Chao Chen, Thanos Bismpigiannis, Radek Grzeszczuk, Kari Pulli, and Bernd Girod, "Outdoors augmented reality on mobile phone using loxel-based visual feature organization," in MIR '08: Proceeding of the 1st ACM international conference on Multimedia information retrieval, New York, NY, USA, 2008, pp. 427–434, ACM

[5] Wei-Chao Chen, Yingen Xiong, Jiang Gao, Natasha Gelfand, and Radek Grzeszczuk, "Efficient extraction of robust image features on mobile devices," in ISMAR '07: Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, Washington, DC, USA, 2007, pp. 1–2, IEEE Computer Society.

[6] R. Szeliski, "Image alignment and stitching: A tutorial", Technical Report MSR-TR-2004-92, Microsoft Research, 2004.

[7] Mann, Steve; Picard, R. W. (November 13–16, 1994). "Virtual bellows: constructing high-quality images from video". Proceedings of the IEEE First International Conference on Image Processing. IEEE International Conference. Austin, Texas: IEEE

[8] Ward, Greg (2006). "Hiding seams in high dynamic range panoramas". Proceedings of the 3rd Symposium on Applied Perception in Graphics and Visualization. ACM International Conference. 153. ISBN 1-59593-429-4.

[9] Steve Mann. "Compositing Multiple Pictures of the Same Scene", Proceedings of the 46th Annual Imaging Science & Technology Conference, May 9–14, Cambridge, Massachusetts, 1993

[10] S. Mann, C. Manders, and J. Fung, "The Lightspace Change Constraint Equation (LCCE) with practical application to estimation of the projectivity+gain transformation between multiple pictures of the same subject matter" IEEE International Conference on Acoustics, Speech, and Signal Processing, 6–10 April 2003, pp III - 481-4 vol.3

[11] K. Elissa, "Title of paper if known," unpublished.

[12] R. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, 2004.

[13] David Chen, Sam S. Tsai, Vijay Chandrasekhar, Gabriel Takacs, Jatinder Singh, and Bernd Girod, "Robust image retrieval using multiview scalable vocabulary trees," in Visual Communications and Image Processing 2009, Majid Rabbani and Robert L. Stevenson, Eds. 2009, vol. 7257, p.72570V, SPIE.

[14] Freeman, W., Pasztor, E., Carmichael, O.: Learning low-level vision. In: Int. Conf. on Computer Vision. (1999) 1182–1189

[15] Efros, A., Freeman, W.: Image quilting for texture synthesis and transfer. Proceedings of SIGGRAPH 2001 (2001) 341–346

[16] A. Agarwala, M. Dontcheva, M. Agrawala, S. Drucker, A. Colburn, B. Curless, D. Salesin, and M. Cohen. Interactive digital photomontage. ACM Trans. Graph., 23(3):294– 302, 2004.

[17] Top 500 Supercomputing sites, top500.org.

[18] https://developer.nvidia.com/cuda-zone.

[19] NVIDIA CUDA C Programming Guide V.4.2,2012, developer.download.nvidia.com/compute/DevZone/docs/html/C/doc/CUDA_C_Programming_Guide.pdf

[20] J. Nickolls, "The GPU Computing Era", Journal IEEE Micro, Vol.30, No.2, pp.56-69, 2010

[21] M. Brown, D. Lowe, "Automatic Panoramic Image Stitching using Invariant Features", International Journal of Computer Vision, No.74, Vol.1, pp. 59-73, 2007

[22] D. Lowe, "Object recognition from local scaleinvariant features", Proc. of Seventh IEEE Int. Conf. on Computer Vision, Vol.2, pp.1150–1157, 1999 /

[23] H. Bay, A. Ess, T. Tuytelaars, L. van Gool "SURF: Speeded Up Robust Features", Journal Computer Vision and Image Understanding, Vol. 110, No.3, pp. 346-359, 2008

[24] M. Fischler, R. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography", Journal Communications of the ACM, Vol.24, No.6, pp.381–395, 1981

[25] Finlayson, G., Hordley, S., Drew, M.: Removing shadows from images. In: ECCV. (2002) IV:823

[26] Levin, A., Zomet, A., Weiss, Y.: Learning to perceive transparency from the statistics of natural scenes. In: NIPS. Volume 15., The MIT Press (2002)

[27] S. M. Seitz and J. Kim. Multiperspective imaging. IEEE Computer Graphics and Applications, 23(6):16–19, 2003.

[28] A. Roman and H. P. Lensch. Automatic multiperspective images. In Eurographics Symposium on Rendering, pages 83–92, 2006

[29] A. Agarwala, M. Agrawala, M. Cohen, D. Salesin, and R. Szeliski. Photographing long scenes with multi-viewpoint panoramas. ACM Trans. Graph., 25(3):853–861, 2006.

[30] J. Yu and L. McMillan. A framework for multiperspective rendering. In EGSR, pages 61–68, 2004.

[31] A. Zomet, D. Feldman, S. Peleg, and D. Weinshall. Mosaicing new views: the crossed-slits projection. IEEE Trans. Pattern Anal. Mach. Intell., 25(6):741–754, 2003

[32] A. Rav-Acha, G. Engel, and S. Peleg. Minimal aspect distortion (mad) mosaicing of long scenes. Int. J. Comput. Vision, 78(2-3):187–206, 2008.

[33] S. Peleg, B. Rousso, A. Rav-Acha, and A. Zomet. Mosaicing on adaptive manifolds. IEEE Trans. Pattern Anal. Mach. Intell., 22(10):1144–1154, 2000.

[34] Fattal, R., Lischinski, D., Werman, M.: Gradient domain high dynamic range compression. Proceedings of SIGGRAPH 2001 (2002) 249–356

[35] Perez, P., Gangnet, M., Blake, A.: Poisson image editing. SIGGRAPH (2003) 313–318

[36] Ballester, C., Bertalmio, M., Caselles, V., Sapiro, G., Verdera, J.: Filling-in by joint interpolation of vector fields and gray levels. IEEE Trans. Image Processing 10 (2001)

[37] Weiss, Y.: Deriving intrinsic images from image sequences. In: ICCV. (2001) II: 68–75

[38] Tappen, M., Freeman, W., Adelson, E.: Recovering intrinsic images from a single image. In: NIPS. Volume 15., The MIT Press (2002)

[39] R. Laganière, OpenCV 2 Computer Vision Application Programming Cookbook, Packt Publishing, 2011

[40] OpenCV Library https://opencv.org/

[41] www.open-mpi.org

Improving Classification Performance for Minority Classes through the use of Positive-Versus-Negative Ensemble Models

Patricia E.N. Lutu
Department of Computer Science,
University of Pretoria,
Pretoria, South Africa
patricia.lutu@up.ac.za

***Abstract*---Classification is a commonly used modelling method for data mining. A classification model is a predictive model which is used to predict a categorical value, called a class. Ensemble classification modelling involves the creation of several base models and a combination algorithm for the base model predictions. The classification modelling process uses a set of instances called the training data. Each instance consists of values for the predictor variables and a categorical label called the class. A class is called a minority class if it has a much smaller number of training instances compared to the other classes. This results in a low level of correct classification compared to the classification performance for the majority classes. Positive-versus-negative (pVn) classification has been reported in the literature as an ensemble classification method which is applicable to classification modelling for multi-class prediction tasks. The purpose of this paper is to report on experimental results for the performance of a replication method for improving classification performance for minority classes, using pVn classification modelling. The experimental results demonstrate that the classification accuracy for minority classes can be improved through the use of pVn classification models.**

***Keywords*---data mining, classification, ensemble modelling, pVn ensemble, ensemble, class imbalance, minority classes**

I. INTRODUCTION

Classification is a commonly used modelling method for data mining. A classification model is a predictive model which is used to predict a categorical value, called a class, given a set of input values for the predictor variables. The modelling process uses a set of instances called the training data. Each instance consists of values for the predictor variables and a categorical label called the class. A classification ensemble model consists of several classification base models, and a combination algorithm which combines the base model predictions into one final prediction. Ensemble classification of a query instance consists of two steps. In the first step, each base model predicts a class to which the instance belongs. Alternatively, each base model provides a prediction score for each of the classes to which the instance could belong. In the second step, the combination algorithm receives the predictions of the base models as inputs, and outputs the final prediction based on a combination rule [1], [2], [3]. Ensemble classification models are known to provide prediction performance improvements compared to single all-classes-at-once (ACA) also known as *k*-class predictive models [1], [2], [3]. Lutu [4], [5], [6] has proposed positive-versus-negative (pVn) modelling as an ensemble classification method that provides predictive performance gains compared to single *k*-class models and, ensembles which use ACA base models [4], [5], [6], [7].

For classification modelling, a data set for model creation and testing is said to be imbalanced if minority classes are present in the data set. These are classes which have much fewer instances in the data set compared to the other classes. When no special intervention is performed for the minority classes, the classification accuracy will be very low for these classes. Many types of interventions for minority classes have been reported in the literature (e.g. Freund and Schapire [8], Hastie and Tibshrani [9], Weiss [10], Seiffert et. al [11], Wang and Yao [12]). For the research reported in this paper, experiments were conducted to create and test pVn ensemble models for a data set which has two minority classes and three majority classes. The first pVn ensemble consisted of unique base models while the second pVn ensemble consisted of replicated base models for the minority classes. Additionally, one ACA ensemble model was created for purposes of comparison with the pVn ensemble classification performance. The experimental results demonstrated that replication of base models which predict minority classes resulted in improved classification performance for the pVn ensemble. The rest of the paper is organised as follows: Section II provides background on ensemble and pVn classification modelling. Sections III and IV provide a discussion of the experimental studies that were conducted to compare the predictive performance of three types of ensemble models, namely: ACA ensemble, pVn ensemble, and pVn ensemble with replicated base models. Section V concludes the paper.

II. BACKGROUND

A. Class imbalance and minority classes

As stated in the last section, a data set for classification model creation and testing is said to be imbalanced if minority classes are present in the data set. These are classes which have much fewer instances in the data set compared to the other classes. There are many application domains of data mining where imbalanced data sets are used. These include credit card fraud [13], forensic data mining for the prediction of telephone fraud [14] and computer network intrusion detection [15]. For these application domains of data mining, the minority class is the class of interest. The class imbalance problem has received a significant amount of attention by researchers in machine

learning and data mining. Boosting and bootstrap sampling (oversampling) are widely used solutions for the class imbalance problem [10], [11].

Galar et. al [16] have reported four categories of methods that have been proposed in the literatures, to address the class imbalance problem. These categories are: algorithm level, data level, cost-based level, and, ensemble methods. Algorithm level methods modify existing algorithms so that they can give special treatment to the training instances for the minority classes. Data level methods use a pre-processing step where the class distribution is changed by increasing the number of training instances for the minority classes. This is commonly achieved through bootstrap sampling (oversampling) of the minority class instances. Cost-based methods combine algorithm and data level methods and additionally use different misclassification costs for each class in the model training step. Ensemble methods use base model diversity, which is known to increase predictive performance. Additionally, these methods may incorporate one or more of the algorithm level, data level, and cost-based methods into the ensemble approach.

Wang and Yao [12] have observed that there are two categories of imbalanced data sets: 2-class and multi-class data sets. Many solutions have been proposed for handling the two-class imbalanced data set problem. However, there are much fewer solutions for handling the multi-class imbalanced data set problem. Wang and Yao [12] have also observed that most multi-class solutions are based on the decomposition of a *k*-class ($k > 2$) problem into fewer class problems e.g. 2-class problems. One example of this approach is the use of One-Versus-All (OVA) classification ensembles [9], [17].

Given a *k*-class prediction task with classes $c_1,...,c_k$, OVA classification involves the decomposition of the prediction task into *k* prediction tasks. A 2-class base model is then created for each class c_i, to predict the (positive) class c_i or the negative class. The negative class consists of all the other *k*-1 classes that appear in the dataset. Each OVA base model is trained on data for class c_i, plus data for all the other classes re-labelled as the class 'negative'. It is then easy to use oversampling in the pre-processing step, in order to increase the training data for the base models which predict the minority classes. It should be noted at this point that pVn ensemble classification uses a data pre-processing step. When minority classes are present, the class distribution is changed by increasing the number of base model training instances for these classes. This is done using bootstrap sampling (oversampling).

B. pVn ensemble classification

Ensemble classification models may be created using ACA base models. An alternative approach for ensemble modelling is to create base models where each base model predicts a subset of the classes of a *k*-class problem (e.g. [4], [5], [17], [18], [19]). For example, the base models for one-versus-all (OVA) ensembles [4], [17], [19] each predict one of the classes of a multiclass prediction task. Positive-versus-negative (pVn) ensemble classification has been proposed by Lutu [4], [5] as a modification to OVA ensemble classification. For pVn classification, a *k*-class ($k > 3$) prediction task is decomposed into *P* sub-tasks. A pVn base model is then constructed for each sub-task. The design of pVn base models is based on information in the confusion matrix obtained by testing the predictive performance of a single *k*-class model created from the given data set, using a specific classification algorithm. The confusion matrix provides useful information about the ease or difficulty that a *k*-class ACA model has, in distinguishing between the different classes. The rationale behind pVn base model design is to create base models where each model specialises in the prediction of a group of classes that a single ACA model has difficulty (confusion) in separating. A pVn base model is designed to predict a subset of the *k* classes of the prediction task. These classes are called the positive classes. All other classes whose instances appear in the training data set for the pVn base model are called the negative classes and are collectively assigned the class label '*other*'. Lutu [4], [5], [7] has reported the algorithm for designing the pVn base models from a confusion matrix and a confusion graph. A confusion graph consists of nodes (vertices) which represent the classes, and arcs which represent the confusion counts in the confusion matrix. The algorithm uses the confusion graph as input, and then determines the positive and negative classes for each base model.

For purposes of base model creation, pVn classification uses a data pre-processing step in order to select the training data for the base models. If a base model has only majority classes, then all the classes have equal numbers of training data instances, which are obtained through random sampling. If a base model has majority and minority classes, then the training data for the minority classes is obtained through bootstrap sampling (oversampling) in order to make the number of training instances to be equal to those for the majority classes. This approach to base model design and implementation makes the pVn ensemble model a suitable solution for modelling from imbalanced data sets. The main objective of the research reported in this paper was to establish whether pVn classification provides a solution for modelling from imbalanced class data sets.

C. Methods for combining pVn base model predictions

When a test or query instance is presented to a pVn base model it can predict one of the *P* classes ($P < k$) it is designed to predict, or it can predict the class '*other*' to indicate that the instance does not belong to any of the *P* positive classes. The base model predictions need to be combined into a single prediction. Various methods have been reported in the literature (e.g. [3], [20]) for the combination of ACA base model predictions for ensemble classification. Suppose that an ensemble model *M*, which consists of *L* base models $m_1,...,m_L$, can predict the classes $c_1,...,c_k$. Let the probabilistic scores assigned to a query instance $\boldsymbol{x}_q$ by the base models $m_1,...,m_L$ for classes $c_1,...,c_k$, be represented by the matrix *DP* with values $DP(m_i,c_j), i=1,..,L, j=1,..,k$. Kuncheva [3] calls this matrix the decision profile for an ensemble model. Each row of *DP* holds the output scores for one base model. Each column of *DP* holds the scores assigned to one class by each of the base models. Kittler [20] and Kuncheva [3] have presented different rules that can be utilised to combine the $DP(m_i,c_j)$ scores in order to select the class

predicted by the ensemble model. These rules include: the *max rule*, the *min rule*, the *product rule*, and the *sum rule*. Lutu [21] has demonstrated that all the four rules can be used effectively for pVn ensemble classification.

The *sum rule* involves the summation of the scores for each class to obtain the combined class score f_j defined as [3]

$$f_j = \sum_{i=1}^{L} DP\ (m_i, c_j) \qquad (1)$$ and

selecting the class with the largest value of f_j. A variation of the *sum rule* is to compute the arithmetic mean of the scores as

$$f_j = \frac{1}{L} \sum_{i=1}^{L} DP\ (m_i, c_j) \qquad (2)$$

and selecting the class with the largest value of f_j. This variation of the *sum rule* was used for the experiments reported in this paper.

III. EXPERIMENTAL METHODS

Experiments were conducted to study the effect of pVn base model replication on predictive performance. The emphasis was placed on the replication of pVn base models which predict minority classes. This section provides a discussion of the data sets used for the experiments, the methods used for base model design, implementation, and testing, as well as the methods for performance evaluation.

A. Data set for the experiments

Two KDD Cup 1999 data sets from the UCI KDD Archive [22] were used for the experiments. The first data set is provided for training data, and the second data set is provided as test data. The small version of the training data set contains 494,022 instances. This version of the data set was used for base model creation. The test data set consists of 311,029 instances. The training and test data sets have 41 predictor variables, also known as predictive features. These data sets consist of a wide variety of computer network intrusions (attack types) that were simulated for a military computer environment. The training data set has 23 classes (attack types) while the test data set has 40 classes. The test set instances that belong to classes that do not appear in the training data set were removed.

TABLE I: CLASS DISTRIBUTION FOR THE REDUCED 5-CLASS KDD CUP 1999 DATA SETS

Class	**Training data**		**Test data**	
	Number of instances	***Percentage of instances***	***Number of instances***	***Percentage of instances***
NORMAL	35,794	68.9	50,285	71.3
DOS	10,851	20.9	10,023	14.2
PROBE	4,107	7.9	4,166	5.9
R2L	1,126	2.2	5,995	8.5
U2R	52	0.1	70	0.1
TOTAL	**51,930**	**100**	**70,539**	**100**

Shin and Lee [23] have reported studies where the 23 classes were grouped into five categories that were treated as the classes for the prediction task. The classes are: NORMAL, DOS, PROBE, R2L, and U2R. The NORMAL class represents non-attack connections to a computer network. The DOS class are the denial-of service attacks. The PROBE class are the probing connections prior to an attack. R2L (remote-to-local) attacks are unauthorised accesses to the network from a remote machine. The U2R (user-to-root) attacks involve unauthorised access to super-user privileges. These five classes were used for the experiments. Further pre-processing was conducted to balance the distribution of the attack types as recommended by Laskov et. al [24]. The final data sets used for the experiments had 51,930 training instances and 70,539 test instances [4], [5], [7]. The minority classes in the training and test data sets are R2L and U2R. Table I gives the class descriptions and distribution for the reduced training and test data sets.

B. Base model design implementation and testing

For predictive classification modelling, it is standard practice to select the relevant features (predictor variables) for the classification model, prior to model construction. These are the features that are correlated to the predicted (class) variable and are not highly correlated to any other predictor variable. The KDD Cup 1999 data set has 41 features. Thirty-four (34) features are quantitative (numeric) and seven (7) features are qualitative (categorical). Selection of the relevant features for the classification models was done using the decision rule-based method of feature selection reported by Lutu [4] and Lutu and Engelbrecht [25]. This method conducts good quality feature selection through a process of measuring the class-feature correlations, and feature-feature correlations from many different samples of the training data set and then computing the mean values of the correlations. Heuristic search is then conducted to determine the best set of features for the predictive classification task. For the KDD Cup 1999 dataset, 32 out of 41 features were selected for model creation.

In order to design the pVn base models, one ACA (5-class) model was created and used to generate the confusion matrix. Stratified sampling was used to obtain a training set of 4,000 instances and validation data set of 2,000 instances that were used to create a multi-layer perceptron artificial neural network (MLP ANN). IBM SPSS 24 statistical software was used for modelling. Five test sets, each with 350 instances were used to test the ACA (5-class) model in order to obtain the counts for the confusion matrix. The confusion graph of Fig. 1 corresponds to the confusion matrix. The algorithm presented by Lutu [4], [5], [7], which was briefly discussed in Section II-B, was used to generate the three pVn base models MNRU, MNDR and MDPR that are shown in Table II.

Each row of Table II gives the model name, positive, and negative classes for one pVn base model. The reader will recall from Section II, that a pVn base model can predict one of the positive classes or the class 'other'. The training data instances of the negative class instances for each base model are collectively re-labelled with the class label 'other' before the base model is created. For example, Table II shows that the MNDR base model has the positive classes: NORMAL, DOS, R2L, and the negative classes:

PROBE and U2R. This means that the training data instances for the PROBE and U2R classes are relabelled as 'other' before the base model is created. So, the MNDR base model is designed to predict a smaller number of classes than the ACA model. This leads to an increase in predictive performance.

In order to create the ensemble base models, stratified sampling from the training and test data was again used to obtain the data sets for training, validation, and testing of the ensemble models' performance. Training set sizes of 4000 instances were used for the ACA and pVn base models. However, for the pVn base models (MNRU, MNRU2, MNRU3), which included the U2R class, the training set size was limited to 1900. This was done to avoid excessive bootstrap sampling (oversampling). Table III provides the description of the three ensemble models that were created. The ACAx3 ensemble consisted of three 5-class base models, ACA1, ACA2, and ACA3. The pVnx3 ensemble consisted of the three pVn base models, MNRU, MNDR, and MDPR. The pVnx5 ensemble consisted of five pVn base models. All three base models for pVnx3 were used for pVnx5. Additionally, the MNRU base model was replicated, by using two additional base models, MNRU2 and MNRU3, each created from different training data.

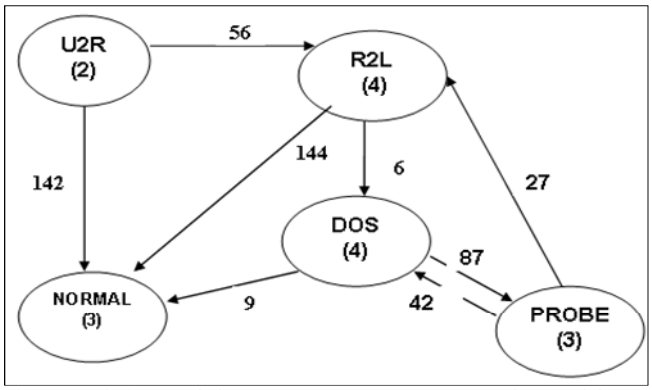


Fig. 1. Confusion graph for KDD Cup 1999 MLP ANN 5-class model

TABLE II: pVn BASE MODELS

Base model name	Positive classes	Negative classes
MNDR	NORMAL, DOS, R2L	PROBE, U2R
MDPR	DOS, PROBE, R2L	NORMAL, U2R
MNRU	NORMAL, R2L, U2R	DOS, PROBE

TABLE III: ENSEMBLE MODELS FOR THE EXPERIMENTS

Ensemble		Base models (positive classes and description)
Name	*Description*	
ACAx3	ACA ensemble	ACA1 (all 5 classes)
		ACA2 (all 5 classes)
		ACA3 (all 5 classes)
pVnx3	pVn ensemble with 3 base models	MNRU (NORMAL, R2L, U2R)
		MNDR (NORMAL, DOS, R2L)
		MDPR (DOS, PROBE, R2L)
pVnx5	pVn ensemble with 5 base models	MNRU (same base model as for pVnx3)
		MNDR (same base model as for pVnx3)
		MDPR (same base model as for pVnx3)
		MNRU2 (MNRU using different train data)
		MNRU3 (MNRU using different train data)

As stated above, the IBM SPSS 24 statistical software was used to create MLP ANN base models. A computer program was implemented in the C++ language for purposes of combining the base model predictions. The version of the *sum rule* given in equation (2) of Section II-C was used for the combination of the base model predictions. In order to measure the performance of the ensemble models, five test data sets of 350 instances (with equal class distribution) for each test set were created through stratified sampling. These test sets were used to test the predictive classification performance of each ensemble model.

C. Methods for performance evaluation

Predictive classification performance was analysed using the analysis of variance (ANOVA) procedure. The analysis of variance (ANOVA) statistical procedure [26] enables the comparison of groups of measurements to determine whether there are statistically significant differences in the performance of the groups. The null hypothesis for the ANOVA procedure is that the population means of the groups are equal. Pairwise comparison of the mean values for many groups requires special treatment. The Scheffé test [26] is appropriate for pairwise comparison of means for experiments involving many groups of measurements. This test was used in addition to the ANOVA procedure in order to establish if there are statistically significant differences in performance between the pVn ensemble with replication (pVnx5) and the other two ensembles (pVnx3 and ACAx3).

Given a test data set with N instances and classes $c_1,...,c_k$. Let the number of test instances of class c_i be denoted by n_i. The testing *accuracy* of a predictive classification model is the percentage of the N test instances whose class labels are correctly predicted by the model. The true positive rate (*TPRATE*) for class c_i is the percentage of the n_i instances whose class labels are correctly predicted as c_i [27]. For the reported experiments, predictive classification performance was measured in terms of mean *accuracy* and mean *TPRATE* for each of the classes. Five test data sets were used to obtain the mean values.

IV. EXPERIMENTAL RESULTS

Table IV provides a summary of predictive classification performance for the three ensemble models. Columns 2, 3 and 4 give the mean *accuracy* for all classes, and the mean *TPRATEs* for individual classes. The measures of *accuracy* and *TPRATEs* are shown as the mean and 95% confidence interval (*CI*) of the mean. The results of columns 2, 3 and 4 indicate that the pVnx5 ensemble model clearly provided the highest mean *accuracy* and mean *TPRATE* for the U2R class.

TABLE IV: PREDICTIVE PERFORMANCE MEASURED ON 5 TEST SETS OF HOLD-OUT DATA

Class	KDD Cup 1999 MLP ANN ensemble models		
	Mean accuracy% or mean TPRATE% and 95% CI of mean for ensemble:		
	ACAx3	*pVnx3*	*pVnx5*
All classes	70.8 ± 2.4	75.0 ± 3.2	78.9 ± 1.8
NORMAL	99.2 ± 0.7	99.7 ± 0.5	99.4 ± 0.5
DOS	66.3 ± 8.5	89.1 ± 9.7	89.7 ± 7.0
PROBE	88.6 ± 2.3	98.6 ± 0.9	98.9 ± 0.7
R2L	56.6 ± 5.7	53.7 ± 4.0	56.3 ± 2.7
U2R	43.5 ± 1.1	33.7 ± 9.4	50.3 ± 3.0

Table V shows the results of the statistical tests to compare the predictive classification performance of the ensemble models in terms of the *accuracy* and class *TPRATE* measures. For purposes of brevity, the three ensembles are represented in columns 2 and 3 as groups *I* and *J*, using the following numbers: 1 for ACAx3, 2 for pVnx3 and 3 for pVnx5. The * character is used to indicate that the mean difference is statistically significant at the 0.05 level. This means that there is a 5% chance of incorrectly rejecting the null hypothesis that the group means are equal.

The Scheffé t-test results of Table V, for *accuracy,* indicate that the pVnx5 ensemble model provided a statistically significant increase of 8.12% compared to the ACAx3 ensemble model. The results for the *TPRATEs* for the DOS and PROBE classes also indicate that the pVnx5 ensemble model provided statistically significant increases of 23.42% for the DOS class, and 10.28% for the PROBE class, compared to the ACAx3 ensemble model. The results for *TPRATEs* for the U2R minority class indicate that the pVnx5 ensemble model provided a statistically significant increase of 16.58% compared to the pVnx3 ensemble model. However, the increase of 6.84% over the ACAx3 ensemble model is not statistically significant at the 0.005 level. This may be due to the fact that only five test sets were used for the reported experiments. Overall, the experimental results provided evidence which indicates that pVn base model replication can lead to increased predictive classification performance for both the minority classes (e.g. U2R) and non-minority classes (e.g. DOS and PROBE). Computer network intrusion data can easily be classified as big data [28]. The characteristics of big data are big volume, big velocity and big variety [28], [29]. Based on the observations from these experiments, the proposed method of pVn ensemble base model replication can be extended to predictive classification modelling from big data.

TABLE V: PAIRWISE COMPARISON OF MEAN ACCURACY AND MEAN TPRATES FOR ACA AND PVN ENSEMBLE MODELS

Perform-ance measure	**Compa-rison between**		**Scheffé t- test**		
	group		***Mean diff. (I-J)***	***95% CI of mean diff.***	***Sig. p-value***
	I	***J***			
Accuracy	3	1	8.12*	[2.65, 13.59]	.005
	3	2	3.94	[-1.53, 9.41]	.176
TPR NORMAL	3	1	0.28	[-.96, 1.56]	.831
	3	2	-0.28	[-1.56, .96]	.831
TPR DOS	3	1	23.42*	[4.46, 42.38]	.016
	3	2	0.58	[-18.38, 19.54]	.996
TPR PROBE	3	1	10.28*	[7.13, 13.43]	.000
	3	2	0.28	[-2.87, 3.43]	.970
TPR R2L	3	1	-0.28	[-9.51, 8.95]	.996
	3	2	2.56	[-6.67, 11.79]	.747
TPR U2R	3	1	6.84	[-5.19, 18.87]	.320
	3	2	16.58*	[4.55, 28.61]	.008

V. CONCLUSIONS

The main objective of the studies reported in this paper was to determine whether the use a pVn ensemble model with base model replication for the minority classes, can provide higher predictive performance for the minority classes. The classification results for the pVn ensemble model with replication were compared with those for an ACA ensemble model, and a pVn ensemble model with no replication. The experimental results reported in Section IV have demonstrated that, for the KDD Cup 1999 data set, the pVn ensemble model with replication provided the highest accuracy (for all classes) and the highest *TPRATE* for the most severe minority class (U2R). These observations indicate that, pVn base model replication has the potential to improve predictive performance for minority classes. For future work, the number of replicated models and the number of data sets for the experiments will be increased in order to obtain more convincing results on the effect of pVn base model replication for minority classes.

REFERENCES

[1] L.K. Hansen and P. Salamon, "Neural network ensembles", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 10, pp. 993-1001, 1990.

[2] L. Breiman, L. "Bagging predictors", Machine Learning vol. 24, pp. 123-140, 1996.

[3] L. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms, John Wiley and sons, Hoboken, New Jersey, 2004.

[4] P. E. N. Lutu, Dataset Selection for Aggregate Model Implementation in Predictive Data Mining, PhD thesis, Department of Computer Science, University of Pretoria, 2010). Available at: http://upetd.up.ac.za/thesis/available/etd-11152010-203041/

[5] P. E. N. Lutu, "Using confusion graphs and confusion matrices to design ensemble base models for classification" in A. Cuzzocrea and U. Dayal (Eds.) Proceedings of the 13th International Conference on Data warehousing and Knowledge Discovery, DaWak 2011, Toulouse, France, August / September 2011, Lecture Notes in Computer Science, LNCS 6862, pp. 301-315, 2011.

[6] P.E.N. Lutu, "Generation of an empirical theory for positive-versus-negative ensemble classification" in Proceedings of the Silver 2012 Workshop/ ECML-PKDD 2012 conference, Bristol, United Kingdom, 24-28 September 2012.

[7] P. E. N. Lutu and A.P. Engelbrecht, "Positive-Versus-Negative classification for model aggregation in predictive data mining", Institute for Operations Research and the Management Sciences Journal on Computing (INFORMS JOC), vol. 25, no. 4, pp. 792-807, 2013.

[8] Y. Freund, Y. and R. Schapire, "Experiments with a new boosting algorithm", Proceedings of the International Conference on Machine Learning, pp. 148-156, 1996.

[9] T. Hastie and R. Tibshirani, "Classification by pairwise coupling", Annals of Statistics, vol. 26, no. 2, pp. 451-471, April 1998.

[10] M. Weiss, "Miningwith rarity: a unifying framework", SIGKDD Explore, vol.6, no. 1, pp. 7-19, 2004.

[11] C. Seiffret, T.M. khoshgoftaar, J. van Hulse, A. Napolitano, "RUSBoost: a hybrid approach to alleviating class imbalance", IEEE Transactions on Man and Cybernetics-Part A: Systems and Humans, vol 40, no. 1, January 2010, pp. 185-197, 2010.

[12] S. Wang and X.Yao, "Multiclass imbalance problem: Analysis and potential solutions", IEEE Transactions on Man and Cybernetics-Part A: Cybernectics, vol 42, no. 4, August 2012, pp. 1119-1130, 2012.

[13] P. Chan and S. Stolfo, "Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection', Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining. AAAI, 1998.

[14] T. Fawcett and F. Provost, F., "Adaptive fraud detection", Data Mining and Knowledge Discovery, vol. 1, pp. 291-316, Kluwer Academic Publishers, Netherlands, 1997.

[15] W. Lee and J. Stolfo, "A framework for constructing features and models for intrusion detection systems", ACM Transactions on Information and System Security, vol. 3, no. 4, 227-261, 2000.

[16] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review of ensembles for the class imbalance problem: bagging, boosting, and hybrid-based approaches", IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews, 2011.

[17] P. E. N. Lutu and A.P. Engelbrecht, "Using OVA modelling to improve classification performance for large datasets" Expert Systems with Applications, vol. 39, no. 4, pp. 4358-4376, 2012.

[18] R. Rifkin and V. Klautau, “In defense of one-vs-all classification”, The Journal of Machine Learning Research vol. 5, pp. 101-141, 2004.
[19] M. Galar, Z. Fernández, E, Barenenchea, H. Bustince and F. Herrara, “An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study of one-vs-one and one-vs-all schemes”, Pattern Recognition, vol. 44, no. pp. 1761-1776.
[20] J. Kittler, “Combining classifiers: A theoretical framework”, Pattern Analysis and Applications, vol. 1, pp. 18-27, 1998.
[21] P.E.N. Lutu, “Empirical comparison of four classifier fusion strategies for positive-versus-negative ensembles”, Proceedings of the South African Institute of Computer Science and Information Technology Conference, SAICSIT 2011. Cape Town, South Africa October 3-5, 2011.
[22] S.D. Bay, D. Kibler, M.J. Pazzani and P. Smyth, “The UCI KDD archive of large data sets for data mining research and experimentation”, ACM SIGKDD, vol. 2, no. 2, pp. 81-85, 2000.
[23] S. W. Shin and C. H. Lee, “Using Attack-Specific Feature Subsets for Network Intrusion Detection”, Proceedings of the 19th Australian Conference on Artificial Intelligence. Hobart, Australia, 2006.
[24] P. Laskov, P. Düssel, C. Schäfer and K. Rieck, “Learning intrusion detection: supervised or unsupervised?”, ICAP: international conference on image analysis and processing. Cagliari, Italy, 2005.
[25] P. E. N. Lutu and A.P. Engelbrecht, “A decision rule-based method for feature selection in predictive data mining”, *Expert Systems with Applications, vol.* 37, no. 1, pp. 602-609, 2010.
[26] P. Cohen, Empirical Methods for Artificial Intelligence. Cambridge, MA:MIT Press, 1995.
[27] T. M. Mitchell, Machine Learning. Burr Ridge, IL:WCB/McGraw-Hill, 1997.
[28] S.M. Othman, F.M. Ba‑Alwi, N.T. Alsohybe and A.Y. Al‑Hashida, “Intrusion detection model using machine learning algorithm on Big Data environment”, *Journal of Big Data, vol. 5, no. 34,* Springer Open, https://doi.org/10.1186/s40537-018-0145-4, 2018.
[29] S. Singh and N. Singh, “Big data analytics”, ICCICT: International Conference on Communication, Information and Computing Technology. Mumbai, India, 19-20 October 2012.

Application of HCI Design Principles in overcoming Information Illiteracy: Case of a M-Health Application for a Rural Community in South Africa

Jeanne Coetzer
Central University of Technology
Bloemfontein, South Africa
jmarkram@cut.ac.za

***Abstract*—Countries in sub-Saharan Africa played a mere spectator role in the first three industrial revolutions. However, researchers, business community, governments and civil society, acknowledges that these countries have the potential to play an active role in the 4th Industrial Revolution. In the case of South Africa, mobile phone technology will play a critical role to this end. It can be argued that the various phases of Information and Communication for Development (ICT4D) have prepared grounds for this shift. However, as articulated in ICT4D 3.0, most ICT4D implementations have fallen trap to the unsustainable and irrelevant transfer of Northern designs to Southern realities. Putting in mind the implicit elements of the status quo, cultural transfers and mutual learning, this paper presents an adaptive, user-centered approach to development of an innovative Mobile-Health (M-Health) intervention for a semi-literate rural community in Sethakeng in the Northern Cape province of South Africa. This is done in the recognition of the fact that one of the main development challenges currently facing the Country is education, skills development and information illiteracy, which is the inability of users to locate, evaluate and effectively use information and communication technology. To achieve the development and implementation of an M-Health intervention, human-computer-interface design concepts of anthropomorphic and affective design principles were employed because they not only lead to mutual-learning but also aspire to implore an emotional or a positively reinforced sub-conscious reaction from users. Three pilot studies carried out using the resulting intervention illustrated a sustained increase in adoption.**

Keywords—ICT4D, 4th Industrial Revolution, user-centered design, human-computer design aspects, affective design, anthropomorphism, emotion, M-Health, adoption

I. INTRODUCTION

Information and Communication Technologies (ICTs) signifies a major opportunity to present positive and substantial lasting changes to the developing world, especially in sub-Saharan Africa, who did not keep pace with these developments [1]. The effective resolving of poverty and human development problems are one of the major challenges facing South Africa as well as many other developing countries [2]. ICTs form an integral aspect of any socio-economic development strategy and if this is not examined, South Africa may just miss out on the Industrial Revolution 4.0. Given the prevalence of mobile phones- even semi-literate and illiterate rural dwellers have mobile phones - these devices can be used to accelerate participation in the Industrial Revolution 4.0. This can only happen if the design of critical applications such as e-health, e-education, etc. is participatory.

ICT for Development (ICT4D) takes the approach of distributing computers, internet and other main stream western technologies to emerging regions without considering the demographics of the intended users and their needs. ICT4D 3.0 puts the user's needs at the forefront and takes a user-centered approach to satisfying these needs [3]. Studies have shown that increased productivity, better health, education and government services result from the extensive adoption of the internet and the countless services and applications that it offers [4]. However, many obstacles remain for the adoption of services provided by the internet especially in rural areas of the developing world [5, 17].

This paper focuses on health and the availability of health services to rural areas described using a case community in the Northern Cape province of South Africa. Although there are many devices and technologies available to modern society to observe and control their health, a gap remains in the provision of such services to rural or urban areas.

The number of individuals using these devices to observe and control their health is now in the hundreds of millions and the rise in the field of M-Health across the world can be traced to the progress of many interconnected trends [6]. According to industry estimations, by the end of 2018, 50% of the more than 3.4 billion smartphone and tablet users will have downloaded mobile health applications [7]. Epidemics and shortage of health workers are continuously presenting challenges to governments and healthcare providers all over the world as determined by the market research company, Statista [8].

Fortunately, the enormous growth of mobile communications and the associated technologies such as smart applications over the last decade offers new hope for the development of quality healthcare [8]. Billions of people now have access to reliable technology, including those who had been left behind in the past by the 'digital divide'.

According to Shah [9], access to healthcare, affordability and quality are problems across the world. There are serious health challenges in developing countries which are further hindered by several core obstacles standing in the way of overcoming this, with the most challenging being the shortage of healthcare workers. Mobile technology opens ways to address these problems and improve healthcare delivery to individuals world-wide. The reach of mobile phones is much wider and further into developing countries than other technology and provides an effective means of bringing health-care services to even the most remote rural areas [10].

Although M-Health applications are mostly distributed on mobile devices such as mobile phones, tablets, and PDAs to provide access to health services and information, they also attempt to affect emotional states [10]. These applications include mobile devices for the gathering of clinical health data in communities, delivery of healthcare information to health practitioners, researchers, and patients. They can monitor patient vital signs and provide care directly to patients.

Mobile devices are becoming increasingly more sophisticated and careful consideration must be given to the design of applications for these devices in terms of their usability, learnability aesthetics, effectiveness, efficiency as well as the emotional aspects of the user experience [11, 18].

An interface that is easy to learn and easy to use is vital for the users, especially novice users. The most important principle of interface design is simplicity that will allow the users to use a product with ease. To achieve simplicity, the design must be consistent and avoid mental overload in the user while at the same time inducing positive emotional responses [12].

Despite the availability of M-Health applications, research has however indicated that in many cases of rural communal settings, Mobile-Health initiatives have not been adopted as expected. Several aspects that contribute to this may include a lack of knowledge regarding the use of the technology utilized as the delivery mechanism, literacy challenges, fear of technology, affordability, and most important; information illiteracy. The "digital divide" is also a factor to consider regarding M-Health initiatives not being adopted. Fortunately, the introduction of ICTs, which included the use of mobile devices, has seen an enormous growth over the past decade, and as such, mobile devices are reaching further into developing countries than any other technology and health infrastructures [13].

The usability and design of these M-Health initiatives may play a crucial role in the adoption and use of such initiatives. Two HCI principles have been identified in this research namely anthropomorphism and affective design.

The assigning of human-like qualities, emotions and thought-patterns to lifeless objects is known as anthropomorphism 14]. Users tend to feel more at ease and experience less apprehension when interacting with and observing non-human objects that have life-like qualities. The second design principle, affective design, describes a design of a product or system that can elicit distinct emotional experiences and feelings from users such as pleasure, confidence, trust, enjoyment, satisfaction [19]. Identifying the emotional connection between users and products with the aim to create products that can elicit pleasure from users is the goal of affective design in Human-Computer Interaction [10].

When applying these design principles to an interface, it may lead to certain outcomes that elicit enjoyment, satisfaction, trust and involvement which will be demonstrated in further sections. Research have demonstrated that the user of a product draws conclusions on the usability of the product based on visual judgments. These studies have demonstrated that users perceive usability as "what is beautiful is useful" [15]. From a usability viewpoint, beautiful products are seen as simple and pleasant to use, which may contribute to a successful and easy to use product [16].

The aim in this research was to explore the impact of the mentioned Human-Computer Interaction (HCI) principles in the overcoming of information illiteracy among users of an M-Health application in a rural community in the Northern Cape province of South Africa.

II. Problem Identified

The incredible pace of dissemination and acceptance of mobile technology and the availability of the internet allows for the provision of many new products, developments and services, especially in health and education. Despite the availability of ICTs across the world and in developing countries, M-Health initiatives in rural communal settings have not been adopted as expected [21]. Factors that contribute to this may include fear of technology and literacy challenges with regards to information illiteracy.

The inability of users to locate, evaluate and effectively use information and technology is known as **"**information illiteracy". Keeping in mind the implied elements of the current situation, cultural transfers and mutual learning, an innovative M-Health intervention for a semi-literate rural community in Sethakeng in the Northern Cape province of South Africa, were introduced whereby a user-centered approach to the development of the intervention was selected.

The researcher aimed to apply the HCI design principles of anthropomorphism and affective design to the existing M-Health application to determine if the appliance of these principles may contribute to the overcoming of information illiteracy among the users in the identified case community while at the same time adopting the usage of the intervention for their health benefits.

The different interface versions of the M-Health application were investigated in terms of enjoyment, satisfaction, trust and involvement that could contribute to the surmounting of information illiteracy of the users.

The research also aimed to implore emotional or positively reinforced subliminal responses from the users when interacting with the M-Health intervention and to determine which design is the preferred version.

III. METHODOLOGY

As previously mentioned, the approach used by ICT4D 3.0 is to include the users in the development process and the best research approach for achieving this was to use a qualitative research design approach, namely Action Research, which is more of a comprehensive approach to problem-solving, instead of a single method for collecting and analysing data [20]. The methods include the keeping of a research journal, observations, recordings, questionnaires, structured and unstructured interviews, as well as case studies. This approach was employed to permit the researcher to attain a deeper understanding of the problem identified. Employing action research as the methodology for the study could also aid in identifying likely difficulties facing rural communities in South Africa [21].

To understand and explore the connection between the case community and the intervention, a case study was used. Three pilot studies were carried out to observe the responses of the case community to the intervention; it allowed the researcher to document and improve the different versions and ultimately describe which design principle is the most effective when applied to the M-Health intervention.

To collect the data, the following data collection methods were employed, which are common to the qualitative research model:

a) Observation: The methodical logging of the case groups' interaction with the application.

b) Questionnaires: Open ended questionnaires to allow respondents to describe their understanding and experience with the application.

c) Focus Groups/Group Discussions: Informal and voluntary gathering with the researcher to discuss ideas, information, proposals and needs regarding the application.

Statistics, generated by an online cloud-based service provider to which the intervention was connected, were used to help the researcher gain insights into the usage of the intervention regarding each of the interaction designs applied.

A. Data Collection Utilizing Action Research

Action research was applied during the project because the interface of the intervention had to be amended a number of times and that these designs had to be re-evaluated through the action research cycles. Action research is described as "learning by doing" and is a process that consists of *examining a problem, taking steps, and discover facts about the result of the steps taken* [12]. The action research process involves four phases or steps namely *plan, act or taking steps, observe and reflect*.

Three design cycles were employed to the user interface of the M-Health application and as such the effects of the different interfaces were investigated during three independent cycles. Therefor action research was utilised over the course of the study. Figure 1 illustrates the intended, inclusive course of action that was applied.

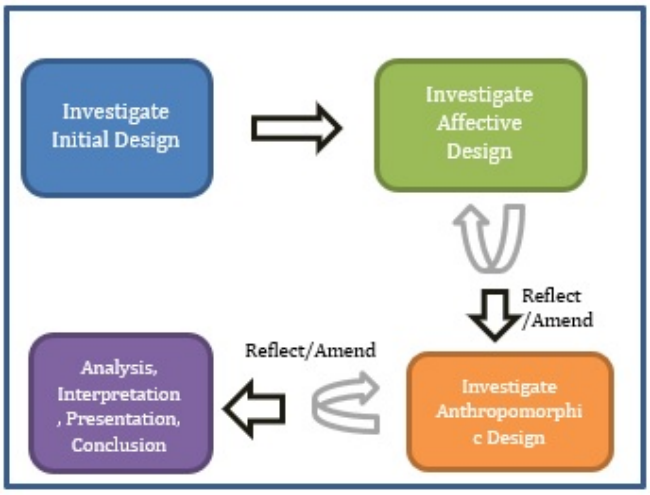


Figure 1. A high-level portrayal of the proposed action research cycles.

B. Data Collection Tools

As mentioned previously, three data collecting tools were utilized namely observations, interviews and focus groups. Proprietary statistical, qualitative data was also gathered from the cloud storage service for gaining insight into the usage of the intervention regarding each of the interaction designs applied. Both quantitative and qualitative data collection techniques were used to gather data with members of the case community playing a vital role in the process.

The total population of the case community at the time of the research was three hundred and ninety three members (393). The sample size was calculated using a confidence interval of 12 and a confidence level of 95%, which resulted in a sample size of at least 57 community members. Six groups were formed from the population sample which consisted of members from the community representing age, gender and social standing. Observations were made and recorded for reflecting and acting for each design phase during the interaction with the interface.

Interviews were held with community members in the form of structured and semi-structured interviews to assist the researcher to gain insight into the understandings and sentiments of the community members concerning the interaction with the technology [21]. Open ended questionnaires were used to collect particular responses from the users which were translated to statistical data referred to in section V. The third data collection tool used was focus groups for the discussion of feelings, attitudes and views of the community members taking part in the data collection regarding the different interface versions presented to them.

This approach is the main focus of ICT4D 3.0 whereby users are to be included in the articulating of their opinions, needs and feelings toward technology. Interactive storyboarding sessions were held and with the aid of infographics, they could articulate their opinions. Figure 2 depicts some of the infographics utilized in these sessions.

Figure 2. The infographics display tools utilized in interactive storyboarding.

The action research steps mentioned earlier were followed and applied to three different cycles and in each cycle a different interface design was investigated [21]. The first cycle was to interact with the first version with which they were already familiar with, the second was based on affective design principles and the third design was constructed and developed using anthropomorphic design principles. The following sections discuss the three versions and the findings in more detail.

IV. Brief Overview Of The Case Community And The Existing M-Health Intervention

This section provides an overview of the case community involved in the study. The information presented is based on reviewed literature and what was gathered through the data collection techniques described earlier.

A rural community in the Northern Cape province of South Africa was identified as a suitable case community in which the research was conducted. The sampling was convenient in nature as a healthcare professional known to the researcher was already working within the community as well as the PhD student who developed the original M-Health intervention and, in turn, knew the researcher. As a result, the process of gaining the community's acceptance and permission to work within it, was less complicated. Sethakeng is situated close to Kuruman and Dithakong in the Northern Cape Province of South Africa. This area of South Africa is also known as the Kalahari.

The first version of the M-Health application related to this study was developed to support the Sethakeng rural community in obtaining meaningful and dependable access to health services and health information. These intentions were founded on the needs of the community as identified by them, as well as those of the healthcare provider mentioned earlier.

Upon investigation, it was found that the general level of literacy was higher than expected but it became evident that the community members struggled with information illiteracy. Although approximately 90% of the community, including adolescents, had mobile devices of different specification levels, many of the community members seemed to have a "fear" of technology.

During the first visit to the community it was observed that the infrastructure of the community was in relatively poor condition, with homes which had very limited access to running water, electricity and other basic services. Roads leading to the community is made of gravel and sand, which make crossing the terrain to connecting communities very challenging. Transport is very limited and only a few community members, other than the king, have motorized vehicles.

Although the community still made use of traditional healers or a Ngak, many seek help from less traditional resources which includes western medicine, health clinics and healthcare providers. The problem facing these communities are the limited access to regular healthcare with the closest town, Kuruman, which is 90 km away. With the challenging road conditions and limited transport, regular access to healthcare services is directly hindered.

It became evident that the Sethakeng rural community experienced several serious challenges and an investigation was launched by the PhD student referred to earlier, to determine if an M-Health application would assist in dealing with some of the healthcare access and delivery needs voiced by the community. The mentioned M-Health application was developed and launched into the community and the re-design of the application's interface was what initiated the basis of this study.

A medical doctor from the nearby town of Kuruman, traveled to the community on certain Sundays of a given month with what can be best described as a mobile health clinic. The doctor provided the community members with basic healthcare services, but the challenge remained for regular services and communication. Therefor the community welcomed the initiative of an M-Health application that could assist them with access to the services of the mentioned medical professional.

The application included the following functionalities:

- Patient-location sharing via GPS to allow for emergency response time;
- Instant messaging (IM) with voice and "chatroom" between linked patients as well as the healthcare professional;
- Image sharing between patient and healthcare professional;
- Remote diagnosis abilities via video-sharing and live video functionality;
- Virtual healthcare notice board to share information with patients and community members in general;
- A scheduler to specify the personal visits of the healthcare professional to the community. Here, the application used visual and audible alerts for easy information distribution.

Over the course of three community visits, the community members were guided to use the application.

V. The Three Interface Versions

A. The Initial Interface Version

The first action research cycle was dedicated to exploring the feelings the community members associated with the use of the initial version of the interface design. The community members were already introduced to the first version by the PhD student who designed the initial application, but this version did not take into account the emotional facets of design as in the case with affective and anthropomorphic design approaches [21]. The goal of the original version was to focus on functionality as a primary design idea without taking HCI design principles into consideration. Data was gathered from the community members taking into account their cultural beliefs, traditional colours and emotional associations that were present within the community. The collected data was applied to re-design the initial version using affective design principles and was passed on to the second action research cycle. Figure 3 depicts an assortment of screenshots of the initial interface version of the M-Health application which indicates the interaction design portion which links the user, the device and the intervention.

The emotional aspects of the user experience, ease of use and simplicity was not taken into consideration with the initial version. With the limited technological literacy of the community taken into account, the initial version did not require a lot of text-based input. This version seemed to be a bit challenging to the community members and referring to the data collected from the first action research cycle, affective design principles were applied to the subsequent version.

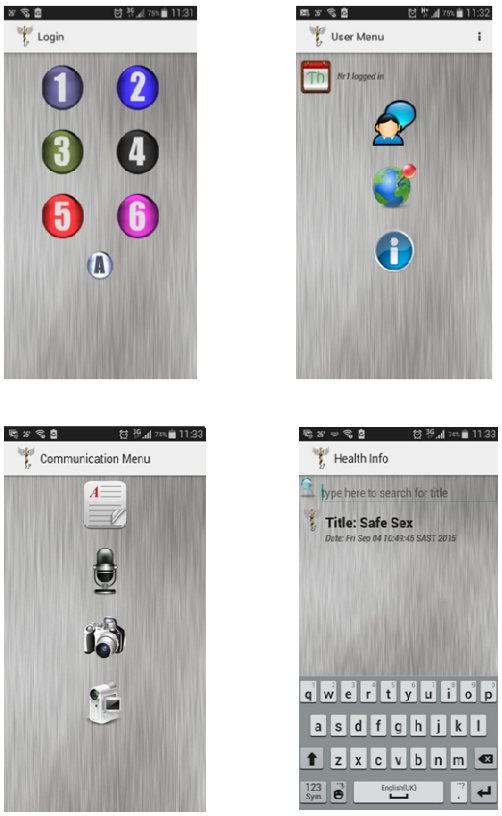


Figure 3. Interface design of the original version of the M-Health Application

B. The Interface Applying Affective Design Principles

The aim of the second research cycle was to examine the effects of an affectively designed interface to determine if such an interface would be more likely to be adopted in terms of usage and to refine the design. The researcher explored colours, images and convictions of the case community to re-design the interface to bring about emotions from the community members. Furthermore, the second round of this action research cycle was used to explore issues that could be useful when considering anthropomorphic interface design within the community.

The information gathered from the literature and research cycle one enabled the researcher to design an interface that aimed to solicit emotional responses form the users and accommodated needs and issues like trust, attitudes and related behaviours. Contextual and cultural aspects were considered for the purpose of eliciting emotional responses from the users. Core emotions like enjoyment, involvement, satisfaction and trust were of particular interest and through the course of the data collection the researcher could establish a link with colours and images held dearly to the community members. These aspects were taken into consideration for the re-design of an affectively design interface and is shown in Figure 4. The functionality of the intervention was not amended but it was founded on the investigation and co-design efforts and input of the community members.

Figure 4: A few interface screens of the affectively re-designed interface.

The goal of the affectively designed interface was to include colours, images and patterns to which the community members could associate and as a result evoke postitive emotional responses from the users. Figure 5 indicates customary colours and patterns that was indicated as being held dear to the community members. During the stage of action research cycle two the investigation focussed on exploring the emotional reactions from the community members toward the affectively designed interface and notes were taken to focus on the comments of the focus groups specifically. Factors taken from this round, was used to re-design the interface in an anthropomorphic manner and then to attempt round three of the action research cycle.

By implementing design principles of affective design, the researcher attempted to afford the user with emotional attractiveness, a positive effect and an overall sense of beauty with regard to the application [8, 14]. This was not fully accomplished in action research cycle two and the final action research cycle was put into motion. The results of the collected data in terms of the core emotions identified

is shown in Figure 9. The only critisism that was received was that, overall, 39% of the interviewees indicated that the representation of the senior and middle-aged group was too western. On the positive side, it was suggested that the affectively designed version was received with more confidence than the original version.

C. The Interface Applying Anthropomorphic Design Principles

Action research cycle three was based on the design principles of anthropomorphism and over the course of action research cycle two, the researcher investigated the possibilty of including virtual agents as a component of anthropomorhic design for the third research cycle. Causing emotional responses while at the same time increase the user experience, were the main goal of action research cycle three.

A mythical character, legendary to the community, was revealed during the data collection process and 25% of the candidates mentioned a legendary healer with magical powers through a doll-like intermediary. The software agents that was designed to represent the fabled characters are depicted in Figure 5. Regrettably 68% of al the interviewees found it rather disturbing and the designer attempted to design an agent with which the community members could feel more comfortable with.

Figure 5. Traditional, fabled characters related to the community

In an attempt to improve the design a talking stethoscope was developed as depicted in Figure 6. "Stetti" was a stethoscope that had could talk and had the same functionality as the talking healer agent. Through the data collected it was however indicated that 64% of candidates had a negative response to these agents. Their feelings conveyed to the researcher were that of mistrust, confusion and not identifying with it.

Figure 6. "Stetti", the interactive agent, failed to prove successful as an elicitor of emotions

Another interview session was held with the community members since the assigning of human-like characteristics to inanimate objects were unsuccessful. During a visit to the community, the named healthcare professional was present, and the researcher observed that the community members held a great deal of admiration and love for the doctor and his accompanying pharmacist. Because of this observation, interactive agents were designed as shown in figure 7, which was a representation of the healthcare provider and the pharmacist.

Figure 7. Representations of the healthcare provider team that showed to be successful as interactive agents

The interactive agents were introduced in the third interface version. During the data collection process, the pshycologist assisting the researcher pointed out that emotional responses of enjoyment and trust among the interviewees were evident. Figure 9 indicates that 72% of the interviewees responded positively to the interactive agent representing the medical doctor and thus the agent was inserted into the anthropomorhic version.

Anthropomorphic principles were included in the re-design of version three of the interface. The needs of the case community as well as the context of the environment were incorporated into the re-design. The third and final interface version of the anthropomorphic design is depicted in Figure 8 showing some of the interface screens.

Figure 8. Anthropomorphically re-designed interface

After investigating the data collected over a three-month period, it was discovered that, in conclusion, the anthropomorphic design was the most effective and desired version. The community members indicated that they felt a sense of ownership with the application because of their constant involvement during the project. The most positive reaction and feedback was from the anthropomorphic design as shown in Figure 9. This fact can be contributed to intense emotions of trust and involvement not felt by the other two designs. Being involved during the design process, community members indicated that they no longer feared the technology and ultimately their information illiteracy were mostly overcome during the course of this project.

VI. Conclusion And Further Work

The ultimate goal of the study was to explore emotional design and how the re-design of an M-Health application's interface would influence the adoption and use of such an application amongst community members within a rural community setting. The main issue that the researcher was faced with, was the level of information and also technical illiteracy among the members of the named case community. In the minds of the community members it was concluded through the collected data, that the anthropomorphic design was superior to the affectively designed interface. The results of the study also illustrated that emotional connections and interface design play a pivotal role in the use and acceptance of an application, although contextually restricted. From reflecting on the data collected in terms of the usage and adoption of the M-Health intervention, it became evident that the community members were more confident in using the technology and contributed to overcoming some of the issues related to information illiteracy. The results were analysed using an inductive approach to data analyis.

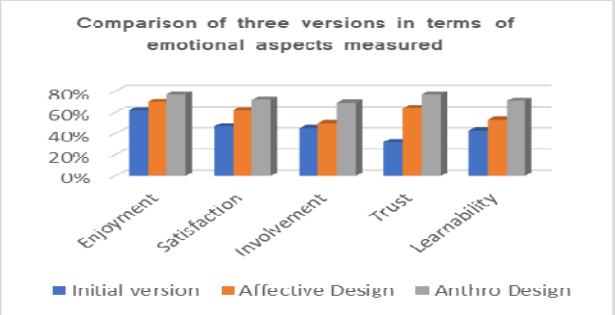


Figure 9. Comparison of the three versions of the M-Health interface using average responses from community members

A. Further Research

- More advanced technology relating to the E-Health context could be included by expanding some of the features of the M-Health application.
- The application could be implemented into a wider environment and investigating the results of such an implementation could prove very discerning.
- Website localization and the impact it has on perceptions and emotions in diverse cultures could be investigated by both researchers and practitioners. Such an investigation could have a huge effect on communities facing similar challenges as the community presented in this study.
- Further work can be done to determine the impact of the chosen design while the M-health application is operational and running.

In conclusion, it can be derived from the research that the role of a well-designed interface plays a pivotal role in strengthening the use of such an intervention. It is also essential to note that a good product does not necessarily mean that it will be used, therefor designers should focus on contextual interface design in the ICT4D environment. This may enhance the use of products, technology and applications especially since the Industrial Revolution 4.0 centers around, among others, ICTs.

REFERENCES

[1] Editor, D.I., The five key challenges in implementing ICT for Development, Retrieved from https/://www.devex.com/new/the-five-key-challenges-in-implmeneting-ict-for-development, 2003.

[2] Fourie, L., Enhancing the livelihoods of the rural poor through ICT: A Knowledge Map. InfoDev, pp. 1-102, June 2008.

[3] Bon, Anna., ICT4D 3.0. – An adaptive, user-centered approach to innovation for development, 2016.

[4] Jeremy Grace & Charles Kenny & Christine Zhen-Wei Qiang & Jia Liu & Taylor Reynolds, "*Information and Communication Technologies and Broad-Based Development : A Partial Review of the Evidence*," World Bank Publications, The World Bank, number 15053, 2009.

[5] Norris, P., The Global Digital Divide: Information poverty and Internet access world-wide, Retrieved from http://w4ra.org, 2006.

[6] Garret, J. J., The Elements of User Experience, New Riders, 2011.

[7] Picard, R. W., Affective computing: challenges. *International Journal of Human-Computer Studies*, pp. 55-64, 2003.

[8] Nace, S., Number of mobile cellular subscriptions per 100 inhabitants in South Afria from 2000 – 2017, Retrieved from https://www.statista.com/statistics/510599/mobile-cellular-subscriptions-per-100-inhabitants-in-south-africa/.

[9] Shah, Anup. "Health Care Around the World." Global Issues, Retrieved from http://www.globalissues.org/article/774/health-care-around-the-world.

[10] Mbipom, G., & Harper, S., Visual Aesthetics and Accessibility: Extend and Overlap. Human Centred Web Lab, School of Computer Science, University of Manchester UK (Unpublished), 2009.

[11] Sanders, E. B., From User-Centered to Participatory Design Approaches, Tayler & Francis Books Ltd., 2002.

[12] Rogers, Y. Sharp, H. & Preece, J., *Interaction Design*. UK: Wiley & Sons Ltd, 2011.

[13] Vital Wave Consulting. *mHealth for Development: The Opportunity of Mobile Technology for Healthcare in the Developing World*. Washington, D.C. and Berkshire, UK: UN Foundation-Vodafone Foundation Partnership, 2009.

[14] McCracken, D. D., & Wolfe, J. R., User-Centered Website Development. Prentice Hall, 2004.

[15] Cyr, D., Emotion and Website Design., Retrieved from Interaction Design Foundation, https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/emotion-and-website-design, 2015.

[16] Tracktinsky, N., Katz, R., & Ikar, D., What is beautiful is usable. *Interacting with Computers*, pp. 127-145, Elsevier, 2000.

[17] The Digital Divide, ICT, and Broadband Internet. Retrieved from Internet World Stats, https://www.internetworldstats.com/links10.htm, August 2018.

[18] Ali, A., Muasaad, A., A study of the interface usability issues of mobile learning applications for smart phones from the user's perspective., *International Journal on Integrating Technology in Education (IJITE)*, vol. 3, no.4, December 2014.

[19] "IGI Global: What is Affective Design?", Retrieved from https://www.igi-global.com/dictionary/design-desirability-collaborative-innovation-initiative/785.

[20] McNiff, J; Whitehead, J., All you need to know about action research. London, UK: Sage. pp. 3–5, 2005.

[21] Coetzer. J, Grobbelaar. L, Masinde. ME. Making software humane. In Proceedings of SAICSIT 2017, Thaba Nchu, South Africa, September 26-28, 2017.

The Effects of a Software Artefact Designed to Stimulate Abstract Thinking Ability on the Academic Performance in Object Oriented Programming of First Year Information Technology Students

Leon Grobbelaar
Department of Information Technology
Central University of Technology
Bloemfontein, South Africa
lgrobbel@cut.ac.za

***Abstract*—A below average throughput of Information Technology students specializing in software development is a challenge that many Universities and Universities of Technology in South Africa face. Contributing factors to this phenomenon are varied at best, but one of the identified factors are that students in this field, especially first year students, find it difficult to conceptualize the associated information and manner of thinking required to become successful in their studies. This is especially true when considering object orientated programming concepts and paradigms that students are required to master as part of their studies. Literary evidence suggests that a high level of working memory, which is associated with abstract thinking ability, is required when learning and applying object orientated programming concepts. The problem becomes more evident and serious if we consider that the Information and Communication Technology sector of a country is largely dependent on the graduating student populous in terms of growing the sector sustainably. A specialized software instrument was developed and tested in an attempt to affect a change in the abstract thinking ability of students from a student sample at a University of Technology. The results of this study focusses on the effect that the instrument realized on the academic performance of first year students related to particularly to object orientated programming and their abstract thinking ability in general as gauged by, amongst other instruments, the General Scholastic Ability Test, or GSAT, rather than focusing on the instrument itself.**

Keywords—Abstract thinking, ICT, ICT4D, Neuroplasticity, Object Orientated Programming, Student performance, Working memory

I. INTRODUCTION

It is commonly accepted, both nationally and internationally, that the Information and Communication Technology (ICT) sector of a country acts as a pivotal driver as far as economic and socio-economic development is concerned. Moreover, The extension of ICT that focuses on the impact that the field on various spectrums of development, commonly referred to as Information and Communication Technology for Development, or ICT4D, has been credited with the power of transformation in developing countries and forms part of many a development agenda of this category of country [1].

South Africa (ZA) is considered a developing country and is subject to also partially rely on the local ICT, particularly the ICT4D industry, when considering growth and development in general [2]. The aforementioned means that the ICT4D industry of ZA, in the guise of both private and public enterprises, largely depend on private and public tertiary institutions like universities and other institutions of higher learning, like the so-called Technical and Vocational Education and Training (TVET) colleges, to provide industry ready, highly trained, apprentice Information Technology (IT) candidates [3].

Public institutions of higher learning as well as private tertiary training facilities are therefore tasked with training and producing students that are skilled enough to help fill the demand for ICT professionals who will ultimately play an important role in ensuring the overall economic growth and socio-economic development in South Africa. This notion is confirmed by the work of Hodge and Miller [4] and a study conducted by the Vaal University of Technology [5] who outlined that education would be one of the South Africa's five medium- to long-term (10 to 20 years) priorities - especially the training and development of skills in the areas of ICT as well as Science and Engineering.

One of the problems that many of the institutions of higher learning that actively offer IT courses face is that of below average student throughput. With an average drop-out rate of approximately 91% between the first- and third-year of study in the field of IT and Engineering, as opposed to the approximate figure of 51% in, for example, Humanities, this is cause for concern [6].

Students arguable face many contextual challenges when it comes to successfully completing their studies, especially first-year students progressing from Matric to this new environment that is the realm of higher education [7]. This inquiry focused on first-year students at a University of Technology (UoT) based in the Free State province of South Africa when it set out to investigate a novel approach to improve the academic performance of first-year IT students particularly delineated to the curriculum dedicated to object-oriented programming (OOP). The investigating party hypothesized that the stimulation and improvement of an individual's abstract thinking ability through the development, implementation and

utilization of a unique software instrument could benefit the academic performance of first-year IT students who are required to understand and apply OOP concepts and paradigms as part of their studies.

Kramer [8] states that, notwithstanding other factors like social and economic challenges, which are beyond the scope of this discussion, a student's performance in OOP-based instructional programs can be directly linked to the manner and level in which a student can apply abstraction, i.e. the level of his or her abstract thinking ability. The related hypothesis was therefore based upon the suggestions made by Kramer and other scholars, as OOP is highly abstract in nature.

A software instrument was developed in association with a trained and registered psychologist who has many years of experience in the GSAT assessment of individuals. This endeavour formed part of a more comprehensive study, but is noteworthy in terms of the contextualization of this particular discussion. The aim of the instrument was to progressively engage and stimulate the part of the human brain associated with a person's abstract thinking ability with the objective to ultimately develop that section of the brain. This notion is based on scholarly diffusion in the sub-field of neuroscience named neuroplasticity. In short, neuroplasticity is the human brain's ability to create new neurons and pathways, or synapses, between the neurons because of certain forms of cognitive stimulation [9]. A first-year IT student sample form a UoT in ZA was used in order to conduct the larger study and investigate the hypothesis.

This paper does not focus on the development of the instrument, the developed instrument itself or the constituents thereof. Rather, in an attempt to align itself with the theme of this conference, which is Technology for Change, this discussion focusses on the change brought about by the use of the instrument within the associated student sample and context. This paper aims to allude to the effect that the use of the instrument affected as far as the following are concerned:

- How did first-year IT students at a UoT in South Africa academically benefit from being exposed to the instrument, if at all?
- How was the abstract thinking ability of the student sample affected by the use of the instrument, if at all?

A shortened literature review is presented in the following section (Section 2), followed by the research design and methodology in Section 3. Thereafter a summarized discussion of the data collection process, data analysis and findings are presented in Section 4. The paper concludes with an overview of the work done and recommendations for future work (Section 5).

II. Review of Literature

Literature points to the fact that many universities and other institutions that offer IT courses that include programming, in particular OOP, as part of the curriculum face the challenge of first-year students struggling to conceptualize the OOP manner of thinking and the related concepts thereof [10]. However, these challenges are not limited to UoTs only. Hernandez [11] quotes Pereira Jr. and Rapkiewickz in their assertion that students at Brazilian Universities who study IT and Computer Science courses encounter difficulties in adapting to the abstraction process when encountering elaborate algorithms on a regular basis, which leads to high failure and dropout rates.

The scholarly contributions that was consulted for this and the larger study seem to suggest the South African institutions of higher and further learning primarily concentrate on teaching the programming languages and how to program, for example teaching the syntax of the programming languages and how to use the integrated development environments (IDE) in the first year of study. This is a necessary and fundamental part of the process as far as teaching computer application programming is concerned. An argument can however be made that by focusing on the foregoing, rather than focusing or at least also addressing how to apply the different fundamentals of programming in varying scenarios and solving different problems with different programming constructs from early on may contribute to students struggling with more advanced and abstract concepts later in their studies [12] [10].

Rogerson and Scott [7] reiterate what was stated the previous section of this deliberation and assert that programming requires a high level of abstraction, which demands both practice and intensive effort to develop the associated parts of the brain responsible for the abstraction process. Abstraction is a mechanism that allows us to represent a complex reality in terms of a simplified model in order to suppress irrelevant details as to enhance comprehension [13]. According to Kramer [8], formal modelling and progressive analysis have proven effective methods in terms of practicing and developing abstract thinking and consolidating the related cognitive ability to apply abstraction.

In their work, Tomlinson et al. confirms the work done by cognitive psychologist, B.F. Pennington, by concluding that working memory is a "computational arena" in which information relevant to a particular task is maintained and subjected to further processing [14]. Working memory therefore plays an important role when it comes to knowledge acquisition through abstraction as the brain continuously receives sensory raw data before editing out data it deems useless. Our ability to represent ideas with symbols and recognize the inherent groupings of objects, amongst other things, also form part of the cognitive abilities associated with abstract thinking and are initially formed and developed at an early age [15] [16].

Piaget outlines the stage of a persons' cognitive development of the brain that is associated with abstract thinking ability within his four stages of cognitive development model [17]. For the sake of brevity, each stage is summarized and outlined as follows:

TABLE I. PIAGET'S FOUR STAGES OF COGNITIVE DEVELOPMENT

Stage 1: Sensorimotor	Stage 2: Preoperational
Characteristics: - Sensory development - Manipulation of objects - Attach names and meaning to objects - Self-centered thinking - Development of imagination, language and memory	Characteristics: - Logical thinking is developed - Abstraction conceptualized through relating concret or physical "things" - Problem solving through physical materials e.g. sand, blocks etc. - Language developed further
Stage 3: Concrete Operational	**Stage 4: Formal Operational**
Characteristics: - Remarkable cognitive growth when compared to other stages - Able to attribute symbols to concrete "things" - Able to attribute symbols to meaning - Logical thinking development - Abstract thinking development - Problem solving development - Critical thinking development - Self-centered thinking reduces	Characteristics: - Make logical use of symbols related to abstract concepts - Abstract thinking patterns are developed vigorously - Symbolic reasoning - Perceptive reasoning is developed - Minor part of self-centered thinking returns

Stages 3 and 4 are the stages in which an individual's abstract thinking abilities are formed. Abstract thinking or thoughts involve a number of different cognitive processes working together to accomplish a particular goal. Tsai and Thomas maintain that abstract thinking evokes schematic processing and focuses primarily on placing information in a larger perspective, to help people understand the bigger picture [18]. Building on Piaget's work, Lister asserted that abstract thinking is the brain's highest level of cognitive functioning that performs crucial and critical processes, e.g. performing calculations and sorting numbers [19]. One can therefore surmise that these thinking abilities enable individuals to use concepts to construct and comprehend generalizations. Moreover, individuals that possess higher levels of abstract thinking ability have the aptitude to categorize, conceptualize and draw conclusions easily from processes that are considered complex [19].

Piaget's work and advances in neuroscience also paved the way for a new theory related to cognitive stimulation and the relation thereof to working memory and abstract thinking ability. In essence, scholars theorized that an individual's working memory could be stimulated in such a manner that the subject's abstract thinking ability could be augmented as to improve the aforementioned.

Klingberg surmised that daily general cognitive processes overload the working memory faculties, which then requires training to function at its full capacity. Furthermore, so-called brain training can improve cognitive performance in a wide range of functions and this improved performance is associated with neuronal changes from the intracellular level to the functional organization of the cortex, known as brain plasticity or neuroplasticity [20]. Working memory can therefore be "exercised" and trained to perform optimally at all times and handle daily workload pressures.

There is also an assertion amongst scholars that, across all cognitive domains examined, subjects engaging in moderate intensity cognitive training programs displayed enhanced cognitive functions [21]. It is noteworthy to mention that the foregoing assertion supports the hypothesis of the larger study related to this discussion that the development of abstract thinking ability is indeed possible and may have a notable effect on the academic performance of students, especially delineated to OOP. This is supported further by the work of Raghubar, Barnes and Hecht who suggest a close relationship between working memory and abstract thinking abilities and the ability of the human brain to function in response to our environmental learning experiences by forming new connections and altering its structure during the human lifespan, referred to as neuroplasticity, as mentioned earlier [22].

Perwej and Parwej brain plasticity or neuroplasticity is the effect of the inherent structure of the brain. The brain can be viewed as a highly complex and sophisticated network of neurons that consists of processing elements called neurons and coefficients (weights), which are bound to connections (synapses), which constitute the neural structure responsible for the training and recall algorithms. This allows the brain to change itself physically and functionally in a number of ways at any age during the learning process. At this stage, indicators show that there are limits to how much the brain can change, which indicates that human abstract thinking abilities can only be stimulated up to a certain level, but further research in this matter is warranted [23].

There are many ways in which to stimulate the cognitive processes related to working memory and, in effect, abstract thinking abilities that fall beyond the scope of this discussion. As the deliberation is based on a software intervention that was developed with the aim to stimulate said cognitive faculties, it is worth mentioning that similar work has been conducted.

In an effort to identify the shortcomings in students' knowledge structures and provide follow-up remedial learning by means of a software resource Chen extended an application of the Pathfinder Network by designing a personalized system for related need [24]. As a result, the Intelligent Tutoring Systems (ITSs) was designed to improve learning in real educational settings. Another example can be found in the work of Alen et al. (2010), who primarily focused on creating a help-seeking ability within their software artefact that provides detailed guidance to students as they learn complex cognitive skills through problem-solving practice [25]. Yet another example where a software intervention was employed to stimulate cognitive processes can be found in the work done by Maillot and Perrot. They used Nintendo Wii's gaming systems such as Wii Sports, Wii Fit and Mario & Sonic's Olympic game, because of their convenience for novice and older populations, their relevance to actual sport, and their ready availability in the marketplace of commercial software to

maximize physiological challenge and increase physical activity. At the conclusion of their work, they found that the use of their platform did improve cognitive performance [26].

Admittedly, a variety of software that focusses on cognitive improvement has been developed in the form of games and so forth [26]. From a uniquely South African perspective though, little has been done to develop contextual software tools that can assist students to stimulate their cognitive processes to enhance their abstract thinking abilities when compared to the international arena. A need for such specialized software does exist, as does the need to consider the criteria to determine how the development and application of these systems can be made suitable for the purpose of learning, whilst addressing the needs of the students directly and indirectly [10].

The literature alluded to in this section supports the notion that working memory and in effect, the abstract thinking ability of an individual can be stimulated through the operationalization of different methodologies with the aim to improve such. The actual result of the method this study prescribed to in regards to this type of stimulation will be discussed in latter sections of this discourse.

III. Research Design and Methodology

A. *Reasoning, Strategy, Unit of Analysis and Sampling*

This study followed a deductive reasoning approach as described by Hofstee [27]. Deductive reasoning was selected as the premise of the study was pre-determined and the conclusions reached were based upon evidence that was supported by the collected and analysed data.

The strategy design was a combined design and consisted of a case study and experiments as outlined by Oliver [28]. The case sampling was convenience (convenience sample) in nature and the sample size was calculated according to the work of Higgins, Bartlett and Kortlik [29]. In short, they describe that the sample size should constitute more than 5% of the sample population at minimum. The sample population for this study consisted of 364 (first year of inquiry) and 348 (second year of inquiry) registered first-year IT students at a UoT in the Free State province of ZA. The sample size related to this study was 80 students per year of inquiry across the two campuses of the UoT involved. Informed consent was appropriated from all participants, including that of the institution, while all ethical considerations as stipulated by the UoT related to this study was observed.

As far as the unit of analysis for this inquiry: This study aimed to investigate the results that the software artefact affected within the case sample as far as academic performance in OOP and abstract thinking ability is concerned. Hence, the unit of analysis was the university assessment results of the participants as well as their performance in the GSAT assessments.

B. *Research Method and Data Sources*

A quantitative approach was employed in terms of the method of inquiry. A quantitative approach was deemed appropriate as the study aimed lay emphasis on the objective measurements and the statistical analysis of the collected data.

The main sources of data were the results of two GSAT assessments and the academic performance records over a two year period of two different samples sizes of 80 students each year over two campuses of the UoT involved in this study. In terms of time horizons related to the data collection; a longitudinal approach was utilized.

IV. Data Collection, Analysis and Findings

A. *Data Collection*

The student samples from the two campuses of the UoT involved in this study was divided into two groups of 20 for each year of inquiry for each campus between the beginning of 2016 and the end of 2017. This totalled a combined sample size of 160 students from a combined sample population of 712 over the period of inquiry.

Two groups from each campus were selected at random and formed the control groups whilst the two remaining groups formed the experimental groups. The groups consisted of volunteer participants of male and female students. The racial demography and the gender distribution of the combined student sample are depicted in Fig. 1 and Fig. 2 respectively with the aim to provide some insight into the composition of the relevant sample.

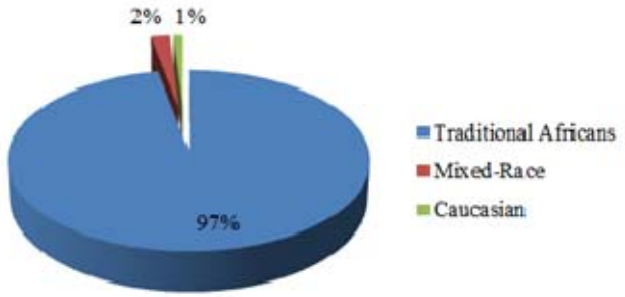


Fig. 1. The racial distribution of the combined student sample.

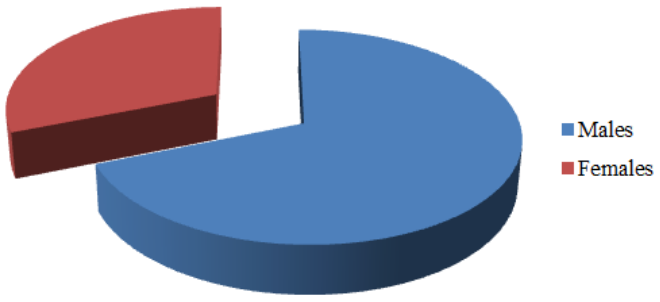


Fig. 2. The gender distribution of the combined student sample.

All the participants from both the experimental and the control groups were subjected to the GSAT assessments. The assessments were conducted and the associated data captured at the inception of each year over the two year period of inquiry. The data was recorded and analysed by the registered psychologist that participated in the study. The aim of subjecting the participants to the GSAT assessments was to scientifically gauge the abstract thinking abilities of the participants and the sample groups at large by assessing non-verbal communication skills. These assessments were referred to as the abstract thinking ability pre-tests.

A copy of the software artefact was installed in a computer laboratory of the UoT at the beginning of 2016. The experimental groups interacted with the software instrument twice a week for 1 hour during the official academic periods of the UoT and were supplied with a copy of the software to install on their personal computers for use during the recess periods. Each experimental group was requested to interact with the instrument in a fashion consistent with that of the official academic period during recess periods. The software was developed to remotely collect usage data and participant performance (with the knowledge and consent of the participants) in the background during participant interaction.

All log files were downloaded from the computers in the computer lab of the UoT and from the personal computers of the participants in cases where the participants did not have access to the internet and the software could not share the data remotely. None of the control groups' participants interacted with the artefact during the period of inquiry.

The academic records of all participants were supplied to the inquiring party by the exam department of the UoT over the period of inquiry and was analysed and aggregated in accordance with the objectives of the study [6]. The motivation behind this endeavour was to collect and analyse tangible empirical data related to the participants' academic performance, especially the data delineated to the participants' performance in subjects that covered OOP and procedural programming. Because an instructional program of procedural programming precedes that of OOP at the particular UoT, the academic data pertaining to student performance in procedural programming was also gathered for comparison purposes. As procedural programming involves less levels of abstraction, it was theorised that a comparison between student performance in procedural programming versus OOP would also be a useful indicator as far as this study is concerned.

Finally, both control- and experimental groups were subjected to the GSAT assessments again at the conclusion of each academic cycle of the UoT over the period of inquiry. This was done to gauge and record any noticeable changes in the abstract thinking abilities of the participants and the sample groups at large and is referred to as the abstract thinking ability post-test. The associated data was again recorded and analysed by the registered psychologist that provided her expertise to the study in that regard.

V. Findings

A. *Abstract Reasoning Ability*

It was necessary to categorize reasoning of the control groups' participants into two sets namely verbal and non-verbal reasoning. This was necessary as to gauge the abstract thinking ability of the sample groups for both the pre- and post-tests as informed by the psychometric (GSAT) assessments.

Non-verbal reasoning abilities represent the abstract thinking or abstract reasoning abilities of the participant groups, whereas verbal reasoning relates the different participant groups' cognitive functioning other than that of abstraction [9]. The psychologist who collected and analyzed the data also included norm descriptions that grouped the verbal and non-verbal reasoning abilities into 5 categories namely *well below average*, *below average*, *average*, *above average* and *well above average*. Fig. 3 graphically represent the reasoning ability distribution of the combined control groups' pre-test.

Fig. 3 relays that 8% of participants scored well below the norm when considering non-verbal reasoning (abstract reasoning). Of the candidates, 28% scored below the average when compared to the norm and 56% of the group tested on par with regard to non-verbal reasoning. In total, 8% of candidates scored above average and 0% well above average for non-verbal reasoning.

In total 36% of the participants scored below the norm in terms of abstract reasoning and only 8% scored above the norm in the same category. Comparatively, a greater number of participants therefore performed below average than above average when considering abstract reasoning ability. Also noticeable is that a greater number of students performed well below average in relation to those who performed well above average.

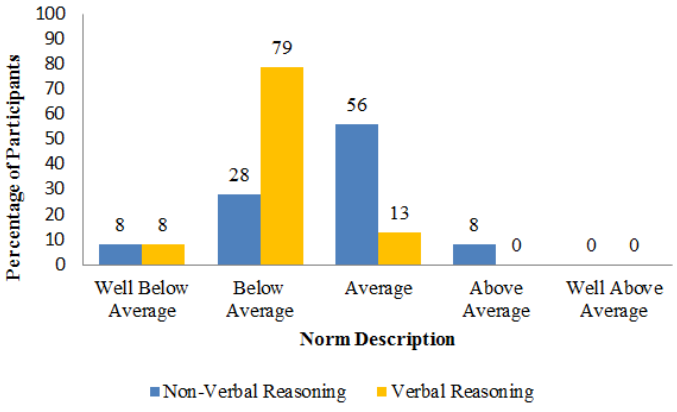


Fig. 3. The verbal and non-verbal reasoning ability of the control groups for the GSAT pre-test.

Fig. 4, on the other hand, represents the verbal and non-verbal reasoning abilities of the combined experimental groups over the period of inquiry.

As Fig. 4 indicates, 43% of the experimental group exhibited below average to well below average abstract

thinking abilities as opposed to the 36% of the participants of the control groups when considering the same category.

In comparison, 0% of the experimental groups' participants exhibiting above average abstract thinking abilities to the 8% of that of the control group. As exhibited in the figure, 0% of the experimental group performed well above average in terms of abstract thinking in relation to the norm.

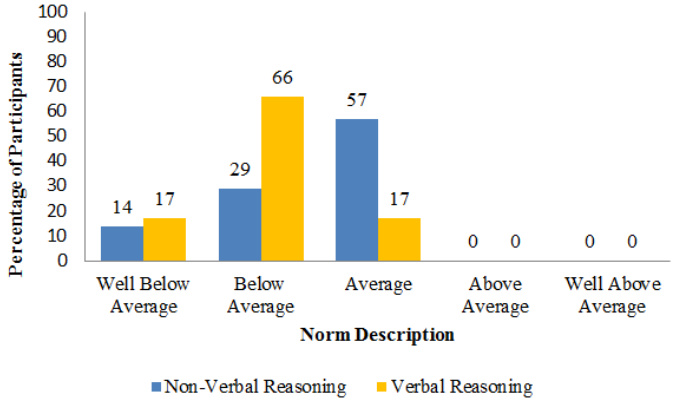


Fig. 4. The verbal and non-verbal reasoning ability of the experimental groups for the GSAT pre-test.

After the abstract thinking ability post-test for the control groups were concluded, the data assimilated and presented in Fig. 5 was extrapolated. The data suggest a 1% decrease of the well below average norm for non-verbal reasoning ability when compared to the pre-test at the time of the post-test's conclusion.

If Fig. 5 is considered, a 1% increase in the below average norm is also noticeable when comparing the pre- and post-test results as well as a 1% increase as far as the above average norm is concerned.

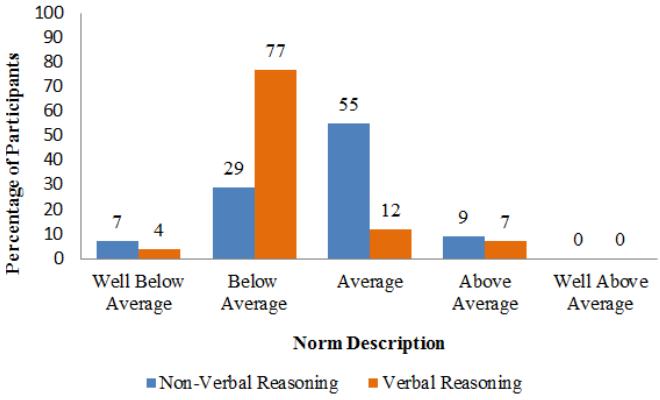


Fig. 5. The verbal and non-verbal reasoning ability of the control groups for the GSAT post-test.

The data of the aggregated experimental groups after the conclusion of the post-test is represented in Fig. 7. A decrease of 8% in the well below average norm of non-verbal reasoning abilities is notable as is the fact that the figure for the below average norm has stayed constant when comparing the results of the pre- and post-tests. What is also clear is that the data shows a 2% increase in the average norm of non-verbal reasoning abilities as well as a 6% increase in the above average norm. This indicates that after the experimental groups engaged the software instrument there was a notable and quite positive shift in the abstract reasoning abilities of the participants of the experimental group.

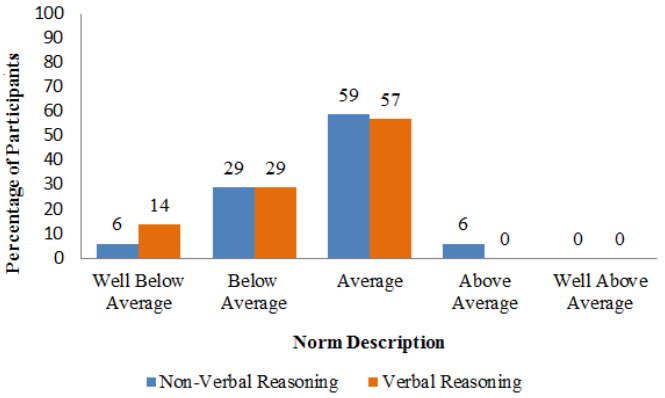


Fig. 6. The verbal and non-verbal reasoning ability of the experimental groups for the GSAT post-test.

B. *Academic Performance*

The academic performance data of the participants of both groups were collected over the period of inquiry. The data consisted of the results of assessments, including assignments and tests, as well as exam papers covering topics related to procedural programming and OOP over the course of the academic period of the UoT. As mentioned in one of the preceding sections, the objective was to employ this data for comparison between the academic performance of the students relevant to the two programming paradigms that is taught at the associated UoT before and after exposure to the artefact.

The data was sorted and combined to represent the academic performance of the participant groups in which participants of both groups received an average of 40% or higher in their overall academic performance regarding each programming paradigm both before the intervention was engaged as well as after the usage cycle of the intervention was concluded. A figure of 40% was selected as this is the minimum that a student must receive to be able to progress academically in the associated institution, albeit that a student that scores between 40% and 49% will have to engage special assessments and achieve 50% to progress to the next level of study. Fig. 7 provides a summarized overview of the data (note that the abbreviation *PP* is used to denote procedural programming).

The data represented in Fig. 7 indicate that the experimental groups experienced a 12% increase in the average performance concerning OOP as opposed to a 2% increase of that of the control groups. The experimental groups' academic performance associated with the procedural programming module declined with 4% when comparing the results, whereas the control groups only experienced a 1% drop in academic performance as far as procedural programming is concerned. It is not clear why the aforementioned decline occurred at this stage.

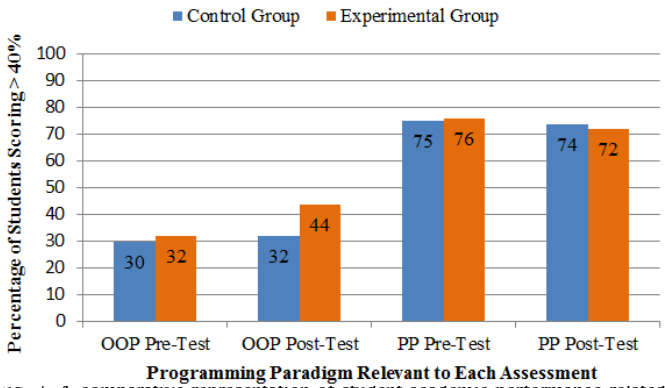


Fig. 7. A comparative representation of student academic performance related to OOP and Procedural Programming before and after the intervention usage cycle was engaged versus completed.

VI. Conclusion and Future Work

This study set out to determine what the effect of a unique, contextually developed software intervention that targeted the stimulation of the cognitive processes associated with working memory would be on the abstract reasoning ability and student academic performance in OOP of first-year IT students at a UoT, if at all. A software instrument of this nature was subsequently developed from the ground up and tested at a UoT as part of a larger study.

This inquiry implemented a case- and experimental research strategy and deductive reasoning to arrive and the concluded results. The main sources of data were the academic records of the participants as well as the results of pre- and post- GSAT assessments.

The data revealed that, in the context of this inquiry, the students who were exposed to the intervention experienced notable positive effects in their abstract reasoning ability. Students using the intervention also displayed an improvement regarding academic performance in OOP when compared to students that did not use the intervention over the period of inquiry.

Although the preliminary data is encouraging, more work needs to be done to assess if the effects achieved are lasting and how the effects can be enhanced further. As mentioned in the literature review section of this deliberation, the neuroplasticity of the human brain reacts to the learning process and known limits, at this stage, on how much the brain can change do exist. The study was limited to one case, therefore, to increase external validity, multiple case studies and a duplication of this inquiry could prove valuable. It may also be of value to make use of advanced brain activity tracking hard- and software like Magnetic Resonance Imaging (fMRI), magnetoencephalography (MEG) or an electroencephalograph (EEG) to map brain activity during the use of the artefact. Also, the notable decline in the academic performance of the participants related to procedural programming after the conclusion of the usage cycle of the intervention could be investigated further to determine the cause and if it is related to the usage of the intervention or not.

The participants of this study did ultimately experienced noteworthy, positive effects in terms of their abstract thinking ability and related academic performance in OOP. More advancements will arguably made in this field of study as neuroscience and IT inform the respective disciplines and pushes the known boundaries of brain development and software tools that may aid the neuro-development process.

Acknowledgment

The author would like to acknowledge and thank Mr. M.S. Kompi for his contribution regarding the larger study associated with this particular inquiry.

References

[1] E. Alampay, "Analysing Socio-Demographic Differences in the Access & Use of ICTs in the Philippines Using the Capability Approach," *The Electronic Journal for Information Systems in Developing Countries,* vol. 27, no. 5, pp. 1 - 39, 2006.

[2] R. Daniels, "Skills Shortages in South Africa: A Literature Review," May 2007. [Online]. Available: https://ssrn.com/abstract=992111 or http://dx.doi.org/10.2139/ssrn.992111. [Accessed 29 June 2018].

[3] A. P. Calitz, J. H. Greyling and M. Cullen, "South African Industry ICT Graduate Skills Requirements," in *SACLA*, Port Elizabeth, 2014.

[4] J. Hodge and J. Miller, "The State-of-the-Art and Implications for National IT Policy," *Information Technology in South Africa,* 1997.

[5] Vaal University of Technology, "Vaal University of Technology," http://www.vut.ac.za/index.php/faculties/applied-a-computer-sciences/departments/information-a-communications-technology 2016. [Online]. Available: http://www.vut.ac.za/index.php/faculties/applied-a-computer-sciences/departments/information-a-communications-technology. [Accessed 28 June 2018].

[6] D. Hlapo, Interviewee, *Mr.*. [Interview]. 28 June 2018.

[7] C. Rogerson and E. Scott, "The Fear Factor: How It Affects Students Learning to Program in a Tertiary Environment," *Journal of Information Technology Education,* vol. 9, no. 2, pp. 148-171, 2010.

[8] J. Kramer, "Is Abstraction The Key To Computing?," *Communication Of The ACM,* vol. 50, no. 4, pp. 37-42, 2007.

[9] L. Bezzola, S. Merillat and L. Jancke, "Motor Training-Induced Neuroplasticity," *GeroPsych,* vol. 25, no. 4, pp. 189-197, 2012.

[10] C. Sanga and M. Venter, "Is a Multi-Criteria Evaluation Tool Reserved for Experts," *Journal Information Systems Evaluation,* vol. 12, no. 2, pp. 165-176, 2009.

[11] C. Hernandez, "Teaching Programming Principles Through a Game Engine," *CLEI Electronic Journal,* vol. 13, no. 2, pp. 1-8, 2010.

[12] M. Armoni, "On Teaching Abstraction in Computer Science to Novice," *Journal of Computers in Mathematics and Science Teaching,* vol. 32, no. 3, pp. 265-284, 2013.

[13] I. Hadar and E. Hadar, "An Iterative Methodology for Teaching Object Oriented Concepts," *Informatics in Education,* vol. 6, no. 1, pp. 67-80, 2007.

[14] R. T. Tomlinson, "Differentiating Instruction in Response to Student Readiness, Interest, and Learning Profile in Academically Diverse Classrooms: A review of literature," *Journal for the Education of the Gifted,* vol. 27, no. 2-3, pp. 119-145, 2003.

[15] F. D. Horowitz, Exploring Developmental Theories: Toward A Structural/Behavioral Model of Development, New York: Psychology Press, 2014.

[16] K. P. Raghubar, M. A. Barnes and S. A. Hecht, "Raghubar, K., Barnes, Working Memory and Working Memory and Mathematics: A Review of Developmental, Individual Difference and Cognitive Approaches,"

Learning and Individual Difference, vol. 20, no. 2, pp. 110-222, 2010.

[17] E. Ojos, "Applying Piaget's Theory of Cognitive Development to Mathematics Instruction," *The Mathematics Educator,* vol. 18, no. 1, pp. 26-30, 2008.

[18] C. I. Tsai and M. Thomas, "When Does Feeling of Fluency Matter? How Abstract and Concrete Thinking Influence Fluency Effects," *Psychological Science,* vol. 22, no. 3, pp. 348-354, 2011.

[19] R. Lister, "Concrete and Other Neo-Piagetian Forms of Reasoning in the Novice Programmer," in *Proceedings of the Thirteenth Australasian Computing Education Conference-Volume*, Sidney, 2011.

[20] T. Klingberg, "Training and Plasticity of Working Memory," *Trends in Cognitive Sciences,* vol. 14, no. 7, pp. 317-324, 2010.

[21] D. Klahr, C. Zimmerman and J. Jirout, "Educational Interventions to Advance Children's Scientific Thinking," *Science,* vol. 333, no. 6045, pp. 971-975, 2011.

[22] K. Raghubar, M. Barnes and S. Hecht, "Working Memory and Mathematic: A review of developmental, individual difference, and cognitive approaches," *Learning and Individual Difference,* vol. 20, pp. 110-122, 2010.

[23] F. Parwej and Y. Parwej, "A Neuroplasticity (Brain Plasticity) Approach to Use in Artificial Neural Network," *International Journal of Scientific & Engineering Research,* vol. 3, no. 6, pp. 1-9, 2012.

[24] L. Chen, "Enhancement of Student Learning Performance Using Personalized Diagnosis and Remedial Learning System," *Computers & Education,* vol. 56, no. 1, pp. 289-299, 2011.

[25] V. Alen, I. Roll, B. McLaren and K. Koedinger, "Automated, Unbostrusive, Action-by-Action Assessment of Self-Regulation During Learning Withe an Intelligent Tutoring System," *Educational Psychologist,* vol. 45, no. 4, pp. 224-233, 2010.

[26] P. Maillot and A. Perrot, "Effects of Interactive Physical-Activity Videogame Training on Physical and Cognitive," *Psychology and Aging,* vol. 27, no. 3, pp. 589-600, 2012.

[27] E. Hofstee, Constructing a Good Dissertation, 1st ed., Sandton: EPE, 2006.

[28] M. S. Oliver, Information Technology Research, Third ed., Pretoria: Van Schaik, 2013.

[29] C. C. Higgins, J. E. Bartlett and J. W. Kotrlik, "Organizational Research: Determining Appropriate Sample Size in Survey Research," *Information Technology, Learning, and Performance Journal,* vol. 19, no. 1, p. 43, 2001.

[30] T. Wegner, Toegepaste Bedryf-Statistiek: Metodes en Toepassings, Cape Town: JUTO en Kie, BPK, 1995.

Evaluation and Modelling of Secured Protocols' Spent Transmission Time

Claude Mukatshung Nawej and Pius Adewale Owolawi
Department of Computer Systems Engineering
Tshwane University of Technology
Pretoria, South Africa
nawejmc@tut.ac.za, owolawipa@tut.ac.za

***Abstract*— In data communication network, one of the ideal network procedures is to transmit files securely from sender to a receiver. However, security alone without speed is not sufficient. In this paper, two network characteristics of an ideal file transfer, security and speed are considered. Using Wireshark network analyzer, we investigate, compare, and finally model the spent transmission time of the same size file between two encrypted WinSCP's protocols, the Secure File Transfer Protocol, and the Secure Copy Protocol. In terms of speed in transferring files, experiment results show that, while retaining the same file size and security integrity, Secure File Transfer Protocol performs faster than its counterpart Secure Copy Protocol does. In order to predict the SFTP protocol speed in transferring files from the slower SCP, a polynomial model of order of three is proposed with a square correlation coefficient of 0.5207.**

Keywords—file transfer, secured transfer, spent transmission time, wireshark

I. INTRODUCTION

One thing computer users cannot do without is the transfer of data over the net. Users make different types of data transfers, unsecured and secure transfers. With the proliferation of internet applications, the security vulnerabilities of data in transit are becoming alarming and these system vulnerabilities are posing a great treat to many gateway servers. Recently, the Federal Bureau of Investigation (FBI) expressed concerns to the health department about impending attacks on anonymous access to FTP servers of medical facilities, which store private health information [1]. File Transfer Protocol (FTP), an intrinsically insecure protocol [2], old but still widely in use, is a communication protocol opened for exchanging files over a Transfer Control Protocol / Internet Protocol (TCP/IP) networks. Therefore, using unsecured protocol of the likes of FTP to transfer data has been classified as unsafe procedures. Sensitive data always require secure transfer protocols. Security is of a big concern in data transfer [3]. However, sufficient speed would be also desirable than just security on its own. In this paper, we compare speed wise the two WinSCP secured protocols, and provide a mathematical equation to model them. The challenge here is to combine security and good speed in file transmission. To do this, we perform a life experiment using WinSCP, and Wireshark.

Wireshark is a popular network analyzer, which is widely used by researchers and network engineers to measure the networks performance and its traffic characteristics. Wireshark has over 500,000 downloads per month [4] which make it the most widely used network analyzer.

WinSCP is a popular file transfer client for Windows, it is often configured to transfer unsecured and secured file. WinSCP is an open source free SFTP client, FTP client, WebDAV client and SCP client for Windows [5]. It was designed to transfer files between local and remote computers/users. Both SCP and SFTP run on SSH (Secure Shell), and are able to provide the same security features. Neither SCP nor SFTP have any file size limitations. The following are WinSCP main features:

WinSCP runs on Windows; It has a graphical user interface; It can upload and download files; It can rename and delete files; It can create files and directories; It synchronizes directories in bidirectional manner; It has both command line console, and a Windows Explorer interface; WinSCP is secured, user friendly, and easy to use. Figure 1 shows a WinSCP login window.

Fig. 1. WinSCP login window

The window login is prompting to provide login information about the location we want to connect to.

The rest of the paper is organized as follows: Section 2 gives a brief review of the work done previously by other researchers. It emphasized gapes needed to be filled up in the current research. This is followed by section 3, the testbed of the experiment. This section deals with the architecture, the topology, and configurations of the scenario. Section 4 compares the two WinSCP's secured protocols with respect to the traffic duration/time, which is one of the distinguished factors considered for the two transfer protocols in this research. Results of how one protocol achieves better performance than its counterpart does are highlighted. In addition, it presents the modelling of the two protocols, where the three-order polynomial is considered. Lastly, section five concludes the research and emphasize the future work.

II. RELATED WORKS

FTP servers are configured to support anonymous login which means that users will be allowed to access files without having to log in remotely to the server. This configuration is done in a way to allow users to connect with their user names "anonymous" which may be known by everyone, and the server itself usually provides automatic password without any form of process or authentication. Because of this insecure log in, an eminent attacker can gracefully connect to this server with the "anonymous" credential, modify or upload files illegally on the server at the expense of the genuine and rightful person.

It is clear that FTP was not designed with strict security measures in mind, nor with an idea of combining security together with speed.

Several researchers have considered the field of transfer protocols, but the reference materials on the combination between protocols types, spent transmission time, and protocols vulnerability all together is limited. In [6] [7] [8] and [9], vulnerabilities in protocols were addressed, these researchers proposed a secured FTP systems. However, some of our other concerns were left aside. In [10], the network speed was considered under different file transfer protocols with less attention to the vulnerability. In their research, Kenneth and Chanson did not put together speed and security. At the contrary, they compared only the performance of some protocols, and derived mathematical models with emphasis on spent transmission speed versus some other parameters of the likes of error rates, and error distributions.

Similar problems were addressed , where the author tackled speed, delay acknowledgement option, and transmission file sizes in [11]. Here, the output of the research concluded that the delay acknowledgement option could improve significantly the protocol speed. In [12], the analysis of speed was related to the hardware in use. In [13], the focus was on the comparison of a secured protocol versus an unsecured one. The research found that the performance of the secured protocol was slower than its counterpart unsecured was.

Going from the abstract above, the argument was that security alone without speed is not sufficient. However, it turns out that fewer studies address or combine secured and fast transmission protocols as an optimum means of file transfer [14] [15]. Therefore, the focus of this paper is to combine security with speed in the transfer of files, compare the performance characteristics of the secured protocols which is speed in our case, and develop a mathematical relationship to model the spent transmission time between the studied protocols.

III. EXPERIMENTAL SETUP

This section deals with the implementation of proposed experiment setup or a test bed so as to accommodate the different scenarios and configurations as expected in the file transfer processes.

A. Testbed Topology

The setup in this paper is a life experiment carried out involving three servers: a Windows application server, a Linux file server, and in addition, a domain controller in a manageable wired network environment. The systems are interconnected using UTP cables, and network switch. The network architecture shown in Figure 2 below is a description of our experimental setup.

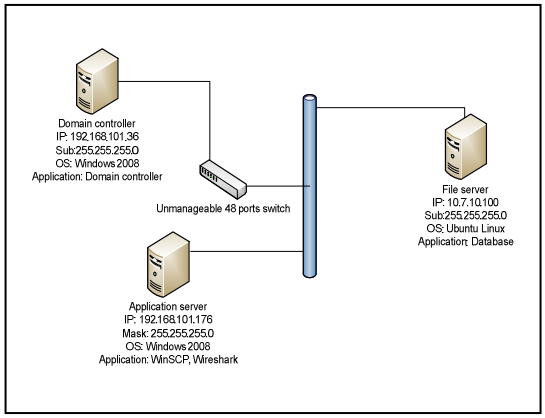


Fig. 1. Hardware setup of the experiment

Figure 2 presents a simple topology of the interconnection of the two actors servers, which are window application server and Linux file server. The Windows application server and the Linux file server are configured as follows:

1) Windows server: This is an application server where Wireshark and WinSCP are running from.

a) Wireshark: The configuration is done as for its network card to capture all traffic passing through.

b) WinSCP: This is a straight forward installation setup which uses its graphical user interface.

2) Ubuntu Linux server: The Ubuntu Linux is the file server where all files are stored for the experiment. The transfer of files are from the Ubuntu file server to the Windows applications server.

3) Third server: The 3rd server is a domain controller, the server in charge of the domain. This server has a role to manage all resources for the entire network.

B. Hardware Specifications

The following is the list and description of hardware devices used in the experiment:

1) File server: CPU: Intel Core i5 @ 2.20 Ghz; OS: Linux (Ubuntu 14); Ram: 64.00 GB; Nic: Realtek PCIe; Database.

2) Application server: CPU: Intel Core i5 @ 3.30 Ghz; OS: Windows server 2008 R2 enterprise full edition; Ram: 2.00 Tb; Application tool for file transfer: WinSCP version 5.7.6; Packet monitor software: Wireshark network protocol analyzer version 2.0.5; Nic: Intel® 82578DM Gigabit network card.

3) Domain controller: CPU: Intel Core i5 @ 3.30 Ghz; OS: Windows server 2008 R2 enterprise full edition; Ram: 2.00 Tb; Nic: Intel® 82578DM Gigabit Network card.

4) Network switch: 48 ports; 10/100/1000Mbps speed.

5) Network cable: Unshielded Twisted Pair (UTP) Cable system; RJ-45 connectors; Cat-5e (Enhanced) category.

C. *Data Transfer*

Figure 3 shows a WinSCP main window. This is a two-part file manager interface. A local folder is displayed on the left pane and a remote folder on the right pane. Files are transferred between these two folders using the mouse or keyboard with standard Windows shortcuts. The experiment used a 2.5 Gb powerIso dragged from the Ubuntu file server, and dropped into the Windows application server using SCP and SFTP protocols alternatively. Follows in Figure 3 below is the WinSCP explorer interface.

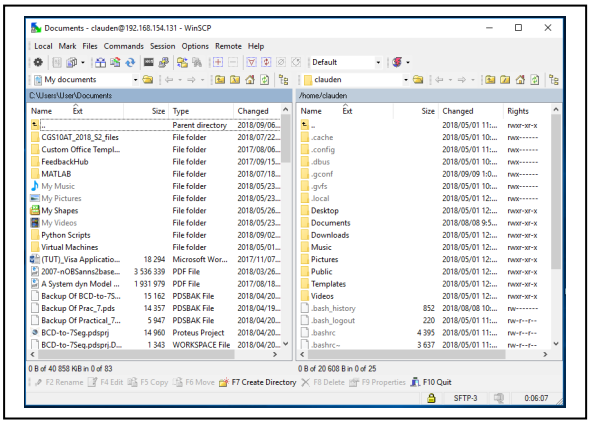


Fig. 2. WinSCP explorer interface

D. *Measurement*

Wireshark has a powerful filter tool, with which it filters the data to reject invalid data and obtain SSH communication data. Time in second is the main performance metrics for the experiment. Figure 4 shows a screenshot of Wireshark capturing messages in transit. This is a live capture process with packet details displayed as they are in transit.

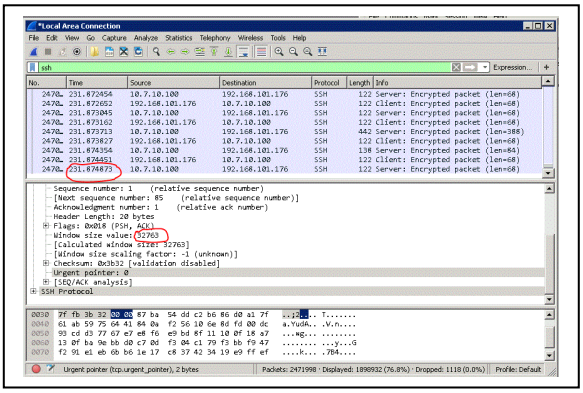


Fig. 3. Wireshark capture

IV. ANALYSIS AND RESULTS

Gauch's theory, the probability of replication [16] states that more measurements of the same experiment leads to greater confidence. In this section, we analyse the transmission speed of the two secured protocols, SCP and SFTP, and present results of the approach employed. To be more accurate, beside from Gauch's theory, we also used the arithmetic means or average calculated value as well as a decider on how many trials to consider for the experiment. Gauch's probability of accuracy giving relationship between number of experimental repeats and confidence of the experiment is summarized in Table I.

TABLE I. PROBABILITY OF ACCURECY IN REPEATING EXPERIMENT (GAUCH, 2006)

Probability of Accuracy in Repeating Experiment	
Number of experimental repeats	***Confidence of success (percent)***
2	60.8
3	66.7
4	70.5
5	73.2
10	80.5
20	86
40	90
100	93.7
162	95
500	97.2

For each protocol, the transmission speed of the same 2.5 Gb file copied from the Ubuntu Linux file server to the Windows application server is measured in turns. We repeat the experiment ten times, and all results averaged over the ten trials.

Table II below lists the calculated spent transmission time of the experiment, which recorded in seconds per trial, and for each protocol. There is significant difference in the spent transmission time between these two protocols.

TABLE II. SPENT TRANSMISSION TIME

Experimental Results		
Number of trials	***SFTP spent transmission time (Seconds)***	***SCP spent transmission time (Seconds)***
1	231.87	291.43
2	236.44	304.43
3	232.71	293.14
4	253.89	305.38
5	236.31	324.45
6	247.58	304.17
7	233.03	299.03
8	231.47	297.07
9	240.13	304.56
10	236.58	305.56

From Table II, we plot Figure 5 below to emphasise the spent transmission time trends difference. We compare the transmission time of the SCP protocol against the one for SFTP protocol using figure 6. The x-axis represents the number of trials (N), and Y-axis represents the spent transmission time (T) for each protocol considered in the experiment.

It is clear that SFTP trend line is different and running lower than the SCP trend line.

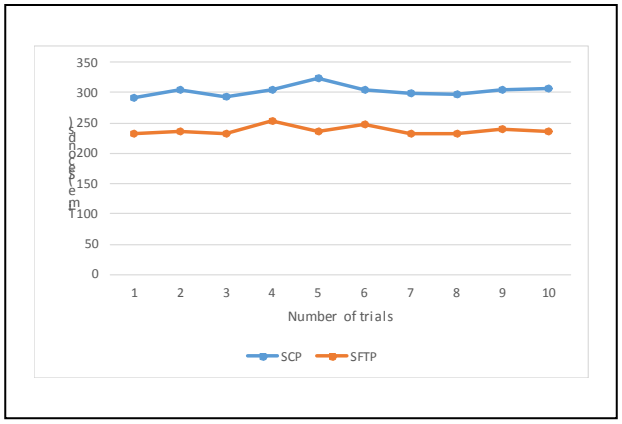


Fig. 4. Spent transmission times for different trials

Considering the lines graph in Figure 5 as the first part of the discussion, in the second part, still comparing the spent transmission time between SFTP protocol and SCP protocol, we calculate the average spent transmission time per protocol as in (1), and use the average values to draw the columns graph in Figure 6. Therefore, emphasizing clearly the difference in size by putting the two columns side by side.

$$A = \frac{1}{n}\sum_{i=1}^{n}(x_i) \quad (1)$$

Where:

A = average (or arithmetic mean)

n = the number of terms

x_i = the value of each individual item in the list of numbers being averaged

Table III gives us the average calculated time of the experiment, which we used to plot column's graph in Figure 6 below to visualise clearly the difference.

TABLE III. AVERAGE CALCULATED SPENT TRANSMISSION TIME

Experimental Results		
	SFTP	***SCP***
Time duration (Seconds)	231.001	302.922

Let us look at Figure 6, in which we compare the average spent time for the two protocols. From this figure, we can see the difference in size between the two columns. Y-axis represents the spent transmission time in seconds, and X-axis represents the two secured protocols of the experiment.

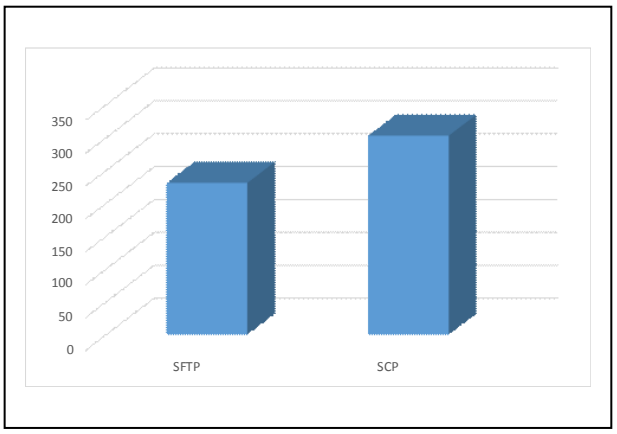


Fig. 5. Comparison of spent transmission time between SFTP and SCP

The bar graph in Figure 6 shows compares the spent transmission time for the two protocols, SCP and SFTP. This reflects a noticeable difference in average spent transmission time between the two considered secured protocols.

The spent transmission time for Secure File Transfer Protocol is smaller than its counterpart Secure Copy Protocol is. Therefore, the Secure File Transfer Protocol runs faster than its counterpart Secure Copy Protocol does.

In addition to the discussion above, a relationship between the SFTP and SCP spent transmission time is developed. A squared correlation coefficient R^2 is used to test the fitness of the model. Considering the best fitting model, a polynomial equation of three order is appropriate with R^2 of 0.5207. Figure 7 below with its x-axis representing the SCP spent transmission time, and the Y-axis corresponding to its equivalent SFTP spent transmission time is based on the derived experimental data from Table II, hence generation of expression for the fitting is presented in equation (2) below:

$$y = 0{,}0055x^3 + 5{,}0674x^2 + 1543{,}4x + 156753 \quad (2)$$

Knowing the behaviour of the SCP protocol, the model equation in (2) provides a convenient prediction tool that can help predict the spent transmission time and behaviour of the SFTP protocol going from the known SCP protocol. Figure 7 shows the prediction pattern between the spent transmission times of the two secured protocols of the experiment.

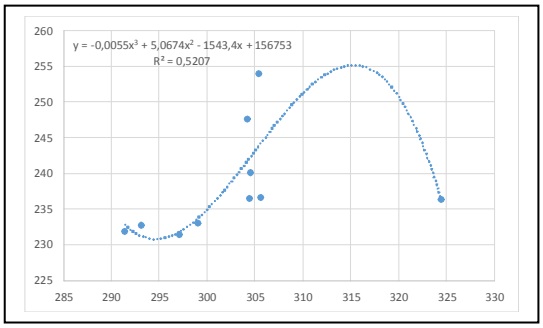


Fig. 6. Predicted values in a polynomial model of order 3

V. CONCLUSION

In this paper, after extensive investigation and comparison of the spent transmission' time between SFTP protocol and SCP protocol, the results in the above experiment are as follows:

'Comparative trend line graph' shows that in all our ten trials, the SFTP spent transmission time is lesser than the spent transmission time in SCP. 'Comparative bar graph for average calculated value' shows that the SFTP spent transmission average calculated spent transmission time is lesser than its correspondent SCP did. In this experiment, with the same file size, SFTP uses less time, therefore, we conclude contrary to previous different statements [17] and research, SFTP protocol is much faster than SCP protocol. A simple polynomial model of three order is statistically significant and appropriate for the relationship between the two secured protocols based on the performance metric of square correlation coefficient of 0.5207. Still in [17], it is assumed that the spent transmission time difference is due to the way these two protocols handle the received packets. In future, extensive work could be carried out, in order to gain more in-deep, analyse, and understand the algorithms which handle the received packets.

REFERENCES

[1] FBI cyber division, private industry notification. (2017, Mar. 22). *Cyber criminals targeting FTP servers to compromise protected health information* [Online]. Available: https://info.publicintelligence.net/FBI-PHI-FTP.pdf

[2] D. Dey, A. Dinda, R. Dinda, P.P. Kundapur, and R. Smitha, "Warezmaster and warezclient an implementation of ftp based R2L attacks," in *Proceedings of IEEE 8th International Conference on Computing, Communication and Networking Technologies*, 2017, pp. 1-6

[3] J. Agarkhed and R. Ashlatha, "An efficient auditing scheme for data storage security in cloud," in *Proceedings of IEEE International Conference on Circuit, Power and Computing Technologies*, 2017, pp. 1-5

[4] L. Chappell, *Wireshark network analysis*, 2nd ed. United States: Laura chappell university, 2010.

[5] WinSCP. (2017). *Introducing WinSCP* [Online]. Available: https://winscp.net/eng/docs/introduction

[6] L. Xia, F. Chao-sheng, Y. Ding, and W. Can, "Design of secure FTP system," in *Proceedings of IEEE International Conference on Communications, Circuits and Systems*, 2010, pp. 270-273

[7] J. Antunes, N. Neves, "Recycling test cases to detect security vulnerabilities," in *Proceedings of IEEE 23rd International Symposium on Software Reliability Engineering*, 2012, pp. 231-240

[8] Y. Ma, H. T. Liu, B. Y. Cai, "Design and implementation of a secure FTP system," in *Proceedings of IEEE Applications and Software Conference*, 2007, pp. 175-176

[9] W. C. He, Y. Y. Zhang, and P. H. Liu. "Research and design of a computer encryption communication system based on secure FTP," in *Proceedings of IEEE Network Security Technology and Application*, 2007, pp. 92-94

[10] K. Lee and S. Chanson, "A new file transfer protocol for telephone lines," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, 1993, pp. 536-548

[11] S. Xu, and T. Saadawi, "Evaluation for TCP with delayed ACK option in wireless multi-hop networks," in *Proceedings of IEEE 54th Vehicular Technology Conference*, 2001, pp. 267-271

[12] R. Raju, F. Scintya, N. Selvaraj, and S. Bhuvaneshwari, "File transfer speed-up by automatic thread assignment in FTP engine," in *Proceedings of IEEE International Conference on Communication and Computational Intelligence*, 2010, pp. 128-131

[13] Y. Wanqing, X. Longteng, and R. Jingyu, "A comparison of TCP and SSL for mobile security," in *Proceedings of IEEE International Conference on Sensor Network Security Technology and Privacy Communication System,* 2013, pp. 206-209

[14] Z. Conghua and C. Meiling, "Analysis of fast and secure protocol based on continuous-time Markov chain," *China communications*, vol. 10, no. 8, pp. 137-149, 2013

[15] C. Rapier and B. Bennett, "High speed bulk data transfer using the SSh protocol," in *Proceedings of 15th ACM Mardi Gras conference*, 2008, pp.11:1-11:7

[16] H.G. Gauch, "Winning the accuracy game," *American scientist* , vol. 94, no.2, pp. 133-141, 2006

[17] JSCAPE LLC. (2018). *Managed file transfer and network solutions* [Online]. Available: https://www.jscape.com/blog/scp-vs-sftp

Graph-Cuts Technique For Melanoma Segmentation Over Different Color Spaces

Olusoji Bidemi Akinrinade & Pius Adewale Owolawi & Chunling Tu
Department of Computer Systems Engineering
Tshwane University of Technology
Pretoria, South Africa
Email: sojiakinrinade2014@gmail.com; owolawipa@tut.ac.za; duc@tut.ac.za

Temitope Mapayi
Department of Electrical Engineering
Mangosuthu University of Technology
Umlazi, Durban, South Africa.
Email: mapayi.temitope@mut.ac.za

***Abstract*—Application of automated image analysis techniques for the detection, diagnosis and management melanoma continues to be an active research area globally. Although a lot of progress has been made on the study of different automated methods of melanoma segmentation, there is still need for further improvement. This paper presents a study on the use of graph-cuts technique for the segmentation of melanoma in clinical images over four different color spaces. The four color spaces considered in this study are RGB, HSV, HSI and HSL. Experimental results show that the use of graph-cuts technique over all the four color spaces are very promising as the average accuracy rate of 96.98%, average sensitivity rate of 89.68%, average specificity rate of 98.96%, average precision rate of 96.34% and average f-score rate of 93,51% are achieved.**

Keywords*—Color Spaces, Graph-Cuts, Melanoma, Segmentation, Skin Cancer.

I. Introduction

Cutaneous melanoma, a cancerous growth in the human skin, continues to be on the increase globally. With an estimated incidence of 62,480 to 178,000 of melanoma cases in the United States between 2008 and 2018, the disease is considered a major deadly form of skin cancer and one of the frequently diagnosed type of skin cancer [1], [2], [3].

Medical imaging, which is a procedure that employs visual diagnostic images of various human body organs for the purpose of medical intervention, continues to find an increasing acceptance in the field of medicine as a major tool for different diseases management [4], [5] . The structures of the various part of human body are captured in the images for the appropriate diagnosis and management of various ailments [4]. Several modalities of imaging have been utilized for the treatment of melanoma and other skin lesions and they can be categorized as dermoscopic and non-dermoscopic images. The non-dermoscopic images are also referred to as clinical images. The use of non-dermoscopic images for melanoma detection has lately been on the increase. The early detection and management of the disease is, however, very imperative because a simple excision can be used to cure it if timely detected. Research on different automated ways of detecting and analysing melanoma using medical images continues to be on the increase globally. Fig. 1 shows some clinical images containing melanoma and manually segmented ground-truth images.

Several automated melanoma detection methods using medical images have been proposed in the literature. In [6] and [7], deep learning methods were implemented for the extraction of lesion in clinical images. The segmentation of melanoma in the clinical images was, however, affected by noise due to non-homogenous illumination and reflection of light across the images. Celebi et al. [8],[9] implemented automated segmentation method based on threshold fusion for the detection of lesion borders in the dermoscopy images. Schaefer et al. [10] combined color normalization technique

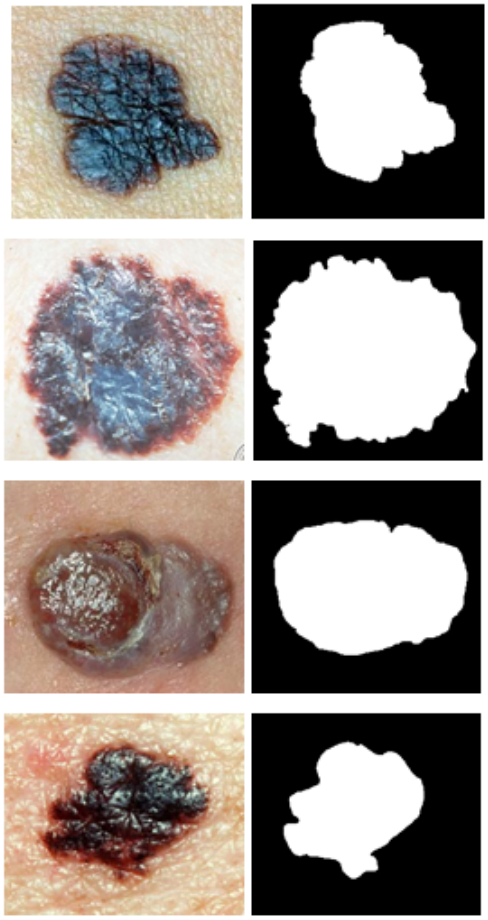


Fig. 1: Melanoma Clinical images & Manually Segmented Ground-Truth Images

and iterative thresholding for the segmentation of melanoma in clinical images.

Although considerable progress has been realized in the literature on the different automated methods of melanoma detection using clinical images, addressing non-homogenous illumination [1], which is one of the major problems of melanoma segmentation in clinical images, needs further study. This paper therefore presents a further study on the use of graph-cuts based technique for melanoma segmentation over different color spaces.

The remaining part of this paper is organized as follows: The methods and techniques utilized in this study are presented in section II. Section III presents the detailed experimental setup utilized in this study with the discussion of the results and section IV draws the conclusion.

II. Methods & Techniques

This section presents the use of graph-cuts technique for the segmentation of melanoma in clinical images. The segmentation approach considered in this study can be divided into three phases namely; pre-processing phase, melanoma Segmentation phase using graph-cuts and the post-processing phase. The study considers four different color spaces for the melanoma segmentation. Although several studies on the segmentation of melanoma in clinical images [1], [9], [11],[12], this study contributes by presenting an investigation on the suitability of different color spaces for melanoma segmentation using graph-cuts segmentation technique.

A. PRE-PROCESSING PHASE

All the melanoma images are in RGB color space. The RGB color space is however different from the human perception of the colors of images [13]. Since one of the major problems of melanoma images is non-homogenous illumination across the images, the melanoma clinical image which is in the RGB color space is converted to HSV (Hue Saturation and Value), HSI (Hue Saturation and Intensity) and HSL (Hue Saturation and Lightness) color spaces. The color space conversion during this pre-processing phase is used to minimize the non-homogeneous illumination problem as the chromaticity color information is separated from the brightness color information in the afore-mentioned color spaces [14], [15], [16], [17].

Given that $R' = R/255, \quad G' = G/255, \quad B' = B/255,$

$$C_{max} = MAX(R', G', B')$$

$$C_{min} = MIN(R', G', B')$$

$\Delta = C_{max} - C_{min}$, the melanoma clinical image is converted from RGB color space to the HSV color space using equation 1.

$$H = \begin{cases} 60^0 \times \dfrac{G' - B'}{\Delta} \ Mod\ 6), C_{max} = R' \\ 60^0 \times \dfrac{B' - R'}{\Delta} + 2), C_{max} = G' \\ 60^0 \times \dfrac{R' - G'}{\Delta} + 4), C_{max} = B' \end{cases}$$

$$S = \begin{cases} 0 \quad , \ \Delta = 0 \\ \dfrac{\Delta}{C_{max}} \quad (\Delta) \neq 0 \end{cases} \tag{1}$$

$$V = C_{max}$$

Another conversion of the melanoma clinical image from RGB color space to the HSI color space is done using equation 2.

$$H = \begin{cases} \Theta \quad , IfB \leq G \\ 360 - \Theta \quad , IfB > G \end{cases}$$

$$S = 1 - \frac{3}{R + G + B} \quad [min(R, G, B] \tag{2}$$

$$I = \frac{1}{3}(R + G + B)$$

Given that Θ is computed as:

$$\Theta = \cos^{-1}\left\{ \frac{\frac{1}{2}[(R - G) + (R - B)]}{[(R - G)^2 + (R - B)(G - B)]^{\frac{1}{2}}} \right\} \tag{3}$$

where R, G and B are the red, green and blue channels respectively of the melanoma clinical image.

The conversion of the melanoma clinical image from RGB color space to the HSL color space is done using equation 4.

$$H = \begin{cases} 60^0 \times \dfrac{G' - B'}{\Delta} \ Mod\ 6), C_{max} = R' \\ 60^0 \times \dfrac{B' - R'}{\Delta} + 2), C_{max} = G' \\ 60^0 \times \dfrac{R' - G'}{\Delta} + 4), C_{max} = B' \end{cases}$$

$$S = \begin{cases} 0 \quad , \ \Delta = 0 \\ \dfrac{\Delta}{1 - |2L - 1|} \quad (\Delta) \neq 0 \end{cases} \tag{4}$$

$$L = \frac{C_{max} + C_{min}}{2}$$

The 'V' channel of HSV color space, 'I' channel of HSI color space and the 'L' channel of the HSL color space are extracted. The extracted image channels and the grayscale image of the RGB color space are individually convolved using a moving median filter of local window size *k*k* as:

$$U(i, j) = H(x, y) * V^1_{k*k}(x, y) \tag{5}$$

where $V^1(x, y)$ (with k=2) represents each of the extracted image channels and the grayscale image of the RGB color space, *U(i,j)* represents each of their respective convolved images and the convolution mask *H(x,y)* is the median filter.

B. Segmentation Phase Using Graph-Cuts

Graph-cuts [18],[19] technique is an energy minimization algorithm extensively applied for image segmentation. It has

been proposed by different authors for different medical image segmentation purposes [20], [21], [22]. It utilizes a weighted graph 'G', with the set of vertices 'V' corresponding to the image pixels 'p' in the clinical image 'P'. The set of edges connecting the vertices is given as 'E', and the two terminal vertices are the sink (i.e. background) represented as "T/0" and source (i.e. object/foreground) represented as "S/1". A set 'N' is used to represent all the neighborhood pairs {p, q} pixels of clinical image 'P' and $L = \{l_1, l_2, ...l_i...l_p\}$ represent the set of labels '0' and '1' (i.e. $l_i \epsilon \{0, 1\}$) that correspond to the background and melanoma respectively in the clinical images.

The best label is assigned to pixel p ϵ P to determine if a pixel is either melanoma (i.e. foreground object) and background in the clinical image. Given that the terminals are separated in the graph, the s/t cut divides the set of clinical image pixels into the two different subclasses. The cut cost 'C' is the sum of the weights of the graph edges that are part of cut. A global minimum s/t cut corresponding to the optimal segmentation of the melanoma in the clinical image amid all the cuts in the graph is then computed in a polynomial time using a Min-Cut/Max-Flow algorithm described by Boykov and Kolmogorov in [19]. Given that R(L) is the regional term, B(L) is the boundary term and α is the relative importance factor between regional and boundary terms, the minimized function corresponding to the optimal segmentation of the melanoma in the clinical image is given as:

$$E(L) = \alpha R(L) + B(L) \tag{6}$$

where the regional and boundary terms are R(L) and B(L) respectively. The regional term R(L) is computed as:

$$R(L) = \sum_{p \epsilon P} D_p(l_p) \tag{7}$$

where the penalty for assigning the pixel p to label l_p is $D_p(l_p)$.

The boundary term B(L) is computed as:

$$B(L) = \sum_{p,q \epsilon N} V(l_p, l_q) \tag{8}$$

such that p and q are neighboring pixels $V(l_p, l_q)$.

The Dp(lp) that evaluates the level to which assigning a label 'fp' to pixel 'p' disagrees with the image pixel intensity (I_p) is computed as:

$$Dp(lp) = -\ln P(I_p|l_p) \tag{9}$$

Given that (p, q) are neighboring pixels $V(l_p, l_q)$, the penalty is 0 when the neighboring pixels have the same labels and this is defined as:

$$V(l_p, l_q) = \begin{cases} 1 & If \quad l_p \neq l_q \\ 0 & If \quad l_p = l_q \end{cases} \tag{10}$$

The sample that shows how the graph that linked the pre-processed clinical image relates to the graph with s/t cut after the clinical image pixels has been divided into background and melanoma subclasses is presented in Fig. 2.

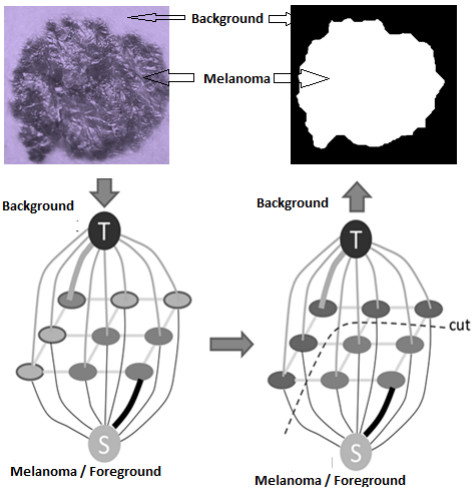


Fig. 2: Graph of Preprocessed Clinical Image and Graph With S/T Cut After Clinical Image is Divided

C. POST-PROCESSING PHASE

A morphological opening based on disk-shaped structuring element is applied on the result obtained to remove the remaining noisy pixels in the segmentation results. The segmentation result obtained from the different color spaces after post-processing is shown in Fig. 3. The clinical images containing melanoma are shown in A1 to A4, the manually segmented ground-truth images in B1 to B4 and the melanoma segmentation results obtained from the grayscale image of RGB color space in C1 to C4. Images D1 to D4 of the Fig. 3 also show the melanoma segmentation results obtained from the HSV color space, images E1 to E4 for the HSI color space and images F1 to F4 for the segmentation results obtained from HSL color space.

III. EXPERIMENTAL RESULTS AND DISCUSSION

Dermatology Information System (DermIS) [23] and DermQuest [24] databases are used in this study. The two databases are used because of their wide usage in dermatology and computer vision research communities. They contain 206 color images and their respective ground-truth images are also provided. 119 of the images are melanoma images and the remaining 87 images are non-melanoma. For the purpose of this study, melanoma images are used for the evaluation of the proposed segmentation method. The performance measures used in this paper are sensitivity, specificity, accuracy, precision and f-score. Given that TP = True Positive, TN = True Negative, FP = False Positive and FN = False Negative. TP is an instance where a pixel is rightly segmented as melanoma while TN is an instance where the the background is correctly segmented. On the other hand, FN is an instance where a pixel belonging to melanoma is wrongly segmented as background while FP is an instance where a background-pixel is wrongly segmented as a melanoma-pixel. The sensitivity measure shows the capability of a segmentation technique to correctly detect the melanoma pixels in the clinical images and the

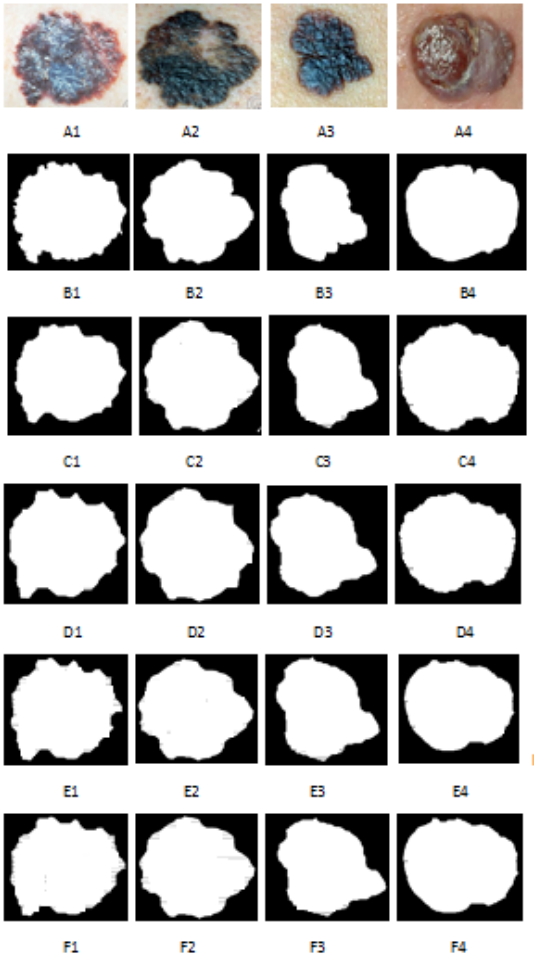


Fig. 3: Color Clinical Images (A1 -A4), Manually Segmented Ground-Truth (B1 -B4) & Melanoma Segmentation Results of RGB color space (C1-C4), HSV color space (D1-D4), HSI color space (E1-E4) and HSL color space (F1-F4)

specificity measure shows the capability of the segmentation technique to correctly detect background pixels. The accuracy measure shows the level to which the segmented melanoma in the clinical image conforms to the ground truth. Precision measure shows the ratio of the correctly predicted melanoma pixels to the total number of pixels predicted as belonging to melanoma in the clinical images. F-score measure computes harmonic average of precision and sensitivity measures. The measures are described in equation (11) equation (15) as follows:

$$Sensitivity = TP/(TP + FN) \tag{11}$$

$$Specificity = TN/(TN + FP) \tag{12}$$

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{13}$$

$$Precision = TP/(TP + FP) \tag{14}$$

TABLE I: Performance of Graph-Cuts Technique For Melanoma Segmentation Over Different Color Spaces

Evaluation Metrics	Color Space			
	RGB	HSV	HSI	HSL
Avg.Accuracy	96.98%	96.44%	96.75%	96.54%
Avg. Sensitivity	88.96%	83.85%	87.74%	89.68%
Avg. Specificity	98.62%	98.96%	98.69%	98.06%
Avg. Precision	91.44%	96.34%	92%	89.30%
Avg. F-Score	93.32%	90.31%	92.70%	.93.51%

$$F - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{15}$$

where Recall = Sensitivity.

The results presented in Table I show the performance obtained from graph-cuts based segmentation of melanoma over the different color spaces. The table shows that the graph-cuts technique perform differently across the different color spaces. The average accuracy rate, average sensitivity rate, average specificity rate, average precision rate and average f-score rate are computed by dividing the total accuracy rate, total sensitivity rate, total specificity rate, total precision rate and total f-score rate respectively by the total number of the segmented melanoma images. The best average accuracy rate of 96.98% is achieved using the grayscale image of the RGB color space and the least average accuracy rate of 96.44% is achieved using HSV color space. Graph-cuts technique over HSL color space achieve the highest average sensitivity rate of 89.68% and HSV color space achieved the least average sensitivity rate of 83.85%. HSV color space, however, achieved the highest average specificity rate of 98.96% and the lowest average specificity rate of 98.06% is achieved using HSL color space. Graph-cuts technique over HSV color space achieved the highest average precision rate of 96.34% and the least average precision rate of 89.30% is achieved when graph-cuts technique is applied over HSL color space. The highest f-score rate of 93.51% is achieved using graph-cuts technique over HSL color space and the least f-score rate of 90.31% is achieved using graph-cuts technique over HSV color space.

IV. Conclusion

This paper presented a study on the use of graph-cuts technique for the segmentation of melanoma in clinical images over four different color spaces (RGB, HSV, HSI and HSL). The experimental results show that the use of graph-cuts technique achieved the average accuracy rates ranging from 96.44% to 96.98%, average sensitivity rates ranging from 83.85% to 89.68%, average specificity rates ranging from 98.06% to 98.96%, average precision rates ranging from 89.30% to 96.34% and average f-score rates ranging from 90.31% to 93,51%. This paper also showed that the investigation of the different color spaces in addressing non-homogenous illumination when combined with graph-cuts technique for the segmentation of melanoma in clinical images look very promising. Our future work shall explore further ways to improve melanoma segmentation performance in clinical images.

Acknowledgement

We would like to thank [23] & [24] for making their dataset publicly available research purpose.

REFERENCES

[1] R. Amelard, J. Glaister, A. Wong, and D. A. Clausi, "High-level intuitive features (hlifs) for intuitive skin lesion description," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 3, pp. 820–831, 2015.

[2] D. S. Cho, S. Haider, R. Amelard, A. Wong, and D. Clausi, "Physiological characterization of skin lesion using non-linear random forest regression model," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*. IEEE, 2014, pp. 3349–3352.

[3] S. Haider, D. Cho, R. Amelard, A. Wong, and D. A. Clausi, "Enhanced classification of malignant melanoma lesions via the integration of physiological features from dermatological photographs," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*. IEEE, 2014, pp. 6455–6458.

[4] K. Doi, "Diagnostic imaging over the last 50 years: research and development in medical imaging science and technology," *Physics in Medicine & Biology*, vol. 51, no. 13, p. R5, 2006.

[5] M. Laal, "Innovation process in medical imaging," *Procedia-Social and Behavioral Sciences*, vol. 81, pp. 60–64, 2013.

[6] N. C. Codella, Q.-B. Nguyen, S. Pankanti, D. Gutman, B. Helba, A. Halpern, and J. R. Smith, "Deep learning ensembles for melanoma recognition in dermoscopy images," *IBM Journal of Research and Development*, vol. 61, no. 4, pp. 5–1, 2017.

[7] M. H. Jafari, N. Karimi, E. Nasr-Esfahani, S. Samavi, S. M. R. Soroushmehr, K. Ward, and K. Najarian, "Skin lesion segmentation in clinical images using deep learning," in *Pattern Recognition (ICPR), 2016 23rd International Conference on*. IEEE, 2016, pp. 337–342.

[8] M. E. Celebi, S. Hwang, H. Iyatomi, and G. Schaefer, "Robust border detection in dermoscopy images using threshold fusion," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*. IEEE, 2010, pp. 2541–2544.

[9] M. Emre Celebi, Q. Wen, S. Hwang, H. Iyatomi, and G. Schaefer, "Lesion border detection in dermoscopy images using ensembles of thresholding methods," *Skin Research and Technology*, vol. 19, no. 1, pp. e252–e258, 2013.

[10] G. Schaefer, M. I. Rajab, M. E. Celebi, and H. Iyatomi, "Skin lesion extraction in dermoscopic images based on colour enhancement and iterative segmentation," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*. IEEE, 2009, pp. 3361–3364.

[11] J. Glaister, R. Amelard, A. Wong, and D. A. Clausi, "Msim: Multistage illumination modeling of dermatological photographs for illumination-corrected skin lesion analysis," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 7, pp. 1873–1883, 2013.

[12] M. Gorriz, A. Carlier, E. Faure, and X. Giro-i Nieto, "Cost-effective active learning for melanoma segmentation," *arXiv preprint arXiv:1711.09168*, 2017.

[13] M. Podpora, G. P. Korbas, and A. Kawala-Janik, "Yuv vs rgb-choosing a color space for human-machine interaction." in *FedCSIS Position Papers*, 2014, pp. 29–34.

[14] A. Ford and A. Roberts, "Colour space conversions," *Westminster University, London*, vol. 1998, pp. 1–31, 1998.

[15] M. Tkalcic and J. F. Tasic, *Colour spaces: perceptual, historical and applicational background*. IEEE, 2003, vol. 1.

[16] B. A. Maxwell, C. A. Smith, and R. M. Friedhoff, "Method and system for separating illumination and reflectance using a log color space," Sep. 29 2009, uS Patent 7,596,266.

[17] B. A. Maxwell, R. M. Friedhoff, and C. A. Smith, "A bi-illuminant dichromatic reflection model for understanding images," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[18] Y. Boykov and G. Funka-Lea, "Graph cuts and efficient nd image segmentation," *International journal of computer vision*, vol. 70, no. 2, pp. 109–131, 2006.

[19] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 9, pp. 1124–1137, 2004.

[20] O. M. Oluyide, J.-R. Tapamo, and S. Viriri, "Automatic lung segmentation based on graph cut using a distance-constrained energy," *IET Computer Vision*, 2018.

[21] J. T. C. Ming, N. M. Noor, O. M. Rijal, R. M. Kassim, and A. Yunus, "Enhanced automatic lung segmentation using graph cut for interstitial lung disease," in *Biomedical Engineering and Sciences (IECBES), 2014 IEEE Conference on*. IEEE, 2014, pp. 17–21.

[22] T. Mashita, J. Usam, H. Shigeta, Y. Kuroda, J. Kikuta, S. Senoo, M. Ishi, H. Matsuda, and H. Takemura, "A segmentation method for bone marrow cavity imaging using graph cuts," in *Pattern Recognition Techniques for Indirect Immunofluorescence Images (I3A), 2014 1st Workshop on*. IEEE, 2014, pp. 20–23.

[23] "Dermatology information system (dermis) database. [online]. available: http://www.dermis.net," Accessed: March 25, 2018.

[24] "Dermquest database. [online]. available: http://www.dermquest.com," Accessed: March 25, 2018.

Asymptotic Analysis of A Fuzzy Based Intrusion Detection System For Zigbee

Tsítso Maphatsoe
Faculty of Engineering and Information Technology
Central University Of Technology
Bloemfontein, South Africa
tsitsomaphatsoe@gmail.com

Muthoni Masinde
Faculty of Engineering and Information Technology
Central University of Technology
Bloemfontein, South Africa
muthonimasinde@gmail.com

***Abstract*—The prospects of achieving a trillion connected internet of things (IoT) devices by 2020 has created the urgency for effective intrusion detection systems (IDS) for these devices. Although it has been argued that the most effective technique used in such systems is anomaly detection, there exist no mechanisms to determine their performance in real-life deployment. In this paper, we report the results of applying asymptotic analysis to evaluate the performance of an anomaly detection algorithm which is designed using logic reasoning through fuzzy logic methodologies. In order to achieve this, the IDS was included as part of intrusion detection software for ZigBee Wireless Sensor Networks (WSNs). In particular, the solution is targeted to address the ZigBee protocol's vulnerability to flood attacks during node discovery and association to the network. The intrusion detection software is hosted external to the WSNs in pursue of a light solution mindful of resource preservation in sensor nodes.**

Keywords—Wireless Sensor Networks, ZigBee protocol, asymptotic analysis, Intrusion Detection Systems, Fuzzy logic, Internet of Things

I. INTRODUCTION

The internet of things (IoT) paradigm is taking its place in business, factories and research laboratories. With this surge, it has been estimated that up to a trillion devices (each with an IP address) will be connected by 2020. This will further complicate the current fragile cybersecurity domain especially now that most of these devices will be vulnerable undocumented first-time creations [1]. The definition of IoT has been elusive, it is no wonder, in [2], it has been described as a 'buzzy phrase' whose interpretation should be based on needs and assets of the proponents. The overreaching concept of IoT is the fact that Wireless Sensor Networks (WSNs) form the baseline technology responsible for basic data collection of the deployment environment. The advances in Micro Electrical Mechanical Sensor (MEMS) engineering and the widespread adoption of wireless communication protocols, for device to device communication, have aided immensely in the popularity of wireless sensor network development and usage [3].

At number 5 of the Forbes Top 10 Tech Trends for 2018, sits "*companies listen in for lucre*" – this refers to the trend where tech companies are eavesdropping on unsuspecting technology users. The most recent example of this is the data breach of over 87 million Face Book users [4]. The ability to carry out this eavesdropping has been inflamed with the rising proliferation of the adoption of smart and intelligent devices – IoT so to say. With the huge developments of IoT in tech-innovations, the opportunities for these kinds of cybercrimes are bound to skyrocket.

By extension therefore, WSNs provide a new door way for malicious software as well as hackers to harm the integrity of data gathered for the numerous IoT applications. This is because they are set up to gather data in inherently insecure environments [5]. It is vital therefore to conclude that WSNs have to remain as secure as possible from any form of intrusion or security breaches. This will bring a sense of acceptance from the end users entering the era of the Internet of Things (IoT), who should also be confident that the information they receive is reliable [6].

It has long been concluded that conventional intrusion detection systems (IDS) such as the ones for Wide Area Networks (WANs) and for Local Area Networks (LANs) are not idea for WSNs. Such IDS should factor the physical limitations of the sensors – they mostly have 32-bit CPUs and 48KB on-board RAM [1]. WSNs are also very sensitive to energy consumption patterns [7]. More critically, the placement of the IDS with the WSN is of paramount importance. For instance, placing it in one of the sensor nodes is extremely vulnerable because of the dynamic nature of these nodes - a node in a WSN can be removed, reprogrammed, and put back into the network [1].

Generally, IDS for WSNs are classified according to the detection techniques utilized; this classification yields three categories: anomalies, state packet examination, or rule-based. In [8], IDS that use the anomaly detection technique and centrally placed, perform better. The big question however still remains; how does one accurately measure this 'performance'? In this paper, we explore the application of asymptotic analysis [9] [10] in assessing the performance of intrusion detection in WSNs that uses the ZigBee protocol for communication. The motivation behind this research is also partially the prospect of developing a WSN security solution as part of a generic integration framework for IoT, which is to be optimized for developing African countries, at the Central University of Technology, Bloemfontein, South Africa [7].

ZigBee is based on the IEEE 802.15.4, and when broken down, its specification can be identified as 802 which is in place for the networking standard; 15 for wireless network and 4 meaning low data transmission rate as well as low power consumption rate [11].

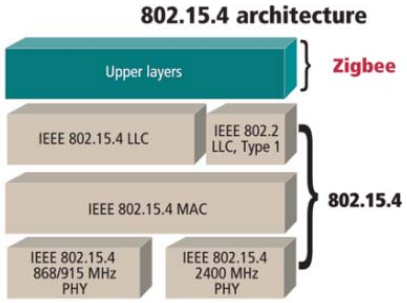


Fig 1: ZigBee Protocol Structure, Source: [11]

ZigBee has been largely developed to be very secure from common intrusion as possible. However, it can be forced into a denial of service state or flood through hacking. This hack is initiated by capturing of a ZigBee node, and re-programming it such that it keeps sending "hello" messages to the central coordinator and re-associating itself anew each time with different address. This is done in an attempt to fill up quickly the database containing the address space of all the nodes that make up the network. Further, the aim is to slow down service to other nodes; such services include usage of the sending and receiving channel and theoretically, only one node will be getting served, the node that is under the control of a hacker [12] [13].

A lot of research has already been done on the need to have Intrusions Detections Systems as a mechanism of ensuring secure networks and the need to build them such that they can be efficient enough to be used in WSNs [9] [8] [14]. This is further amplified by the fact that there are several angles that are being pursued in the development of suitable IDS for each particular network need. For instance, a genetic algorithm approach has been used to develop an IDS and pursuant to the objective, received desired results [15] . It has also been stated that mobile agent based IDS have an advantage over their more static counterparts, such as host based IDS, when it comes to anomaly detection and use of resources [15] [16].

Despite the above, the authors identified a gap in that, the performance of all these IDS has not been scientifically assessed. This led to the study reported in this paper and was guided by the objective of applying asymptotic analysis to measure such performance. In doing so, the Big OH (O) notation was used [10].

II. Algorithm Design and Implementation

A. Intrusion Detection Algorithm

Design and analysis of algorithm has established itself as an important topic in computer science because it makes it possible to evaluate the performance of yet to be implemented software in a language- and machine-independent fashion [17]. Using this approach, the algorithm shown in fig.2 below was analysed and then used to design the IDS. Adopted and adapted from [18], the algorithm objective is to thwart attacks aimed at ZigBee sensors. The algorithm uses a different approach of monitoring the hash tables of the protocol to check on rogue sensor nodes that have changed their mac addresses. The algorithm has two stages, the *learning* stage and the *detection stage*.

```
evaluation_time = get_time() + evaluation_step
loop
  receive_byte()
  if message_complete() then
    evaluate_address()
    update_hashtables() {update short- and long-address
    hashtables}
    if get_time() > evaluation_time then
      if get_environment() + threshold <
      size(hashtable) then
        drop_communication() {intrusion detected: drop
        message}
      else
        update_environment()
        update_thresholds()
      end if
      evaluation_time = get_time() + evaluation_step
    end if
  end if
end loop
```

Fig 2: Security Algorithm for Detecting Flood Attacks in ZigBee WSNs [18]

In this research, we have used the template design model of algorithm design, based on the above mentioned algorithm as a template. In this model, the overall flow of the algorithm is maintained but certain subclasses or modules/sections are modified for a specific purpose such as our approach. In doing so, the concepts of analysis and design of algorithms as described in [19] are followed. For instance, in this research, a database is built into the IDS software to keep track of node addresses and to help the algorithm when it tracks down rogue nodes. Additionally, the evaluation is not done at the protocol hash tables but from the database. The aim is to evaluate a threat before the hash tables crash after being flooded by rogue nodes. These tables keep track of all sensor nodes in the network as well as their associated addresses. A flood, in essence happens when these tables are being forced to record more addresses than they can hold [18]. This happens in the window period afforded by the protocol during discovery, where a node answers a call from the gateway to send in data. A node replies by giving its mac address and node ID during the association stage.

A rogue node can, at this point, keep answering to the association call, with as many different addresses as possible within a short space of time, leading to a rapid filling up of hash tables and eventual crash of the network. Finally, this research takes into consideration that a node can be considered bad based on limited evaluation knowledge and therefore the algorithm is augmented with a fuzzy logic module to lower the possibilities of falsely claiming a node to be under attack. The desktop application software is written in JAVA and is adapted from a Java framework meant for com-port programming. It monitors sensor nodes deployed as though in field of which, in this case, are still in a controlled indoor environment. These sensors are in turn sending their data readings to a computer through the Libelium ZigBee gateway.

B. *Architecture of the Intrusion Detection System Application*

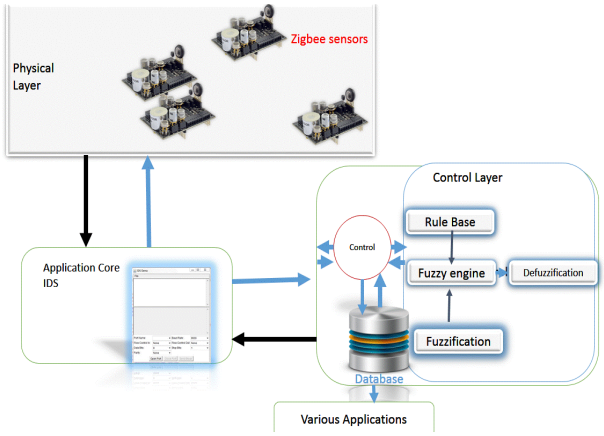


Fig 3: IDS Architecture

1) Physical Layer (Libelium Waspmote ZigBee Sensors)

The physical layer is the point through which data is collected. Sensors transmitting through the ZigBee 802.15.4 communication protocol make up this layer, for the purpose of this research. The ZigBee gateway that collects data from the sensors to the computer is also found at this layer. These sensors have been programmed such that with every data packet frame they send, for whatever sensed data of interest to an end-user, they also broadcast their mac address as well as their node IDs. The sensors are also programed to listen to Over the Air (OTA) commands, as well as to broadcast at only a pre-determined baud rate and Pan ID. This is done because they would not be able to locate the gateway if it is operating on a different Pan ID. Additionally, the IDS software would not be able to decode the data packets if it read them at a different baud rates other than the one at which they are sending data. This would cause problems for the control layer as it would not be able to extract desired information from the packet.

2) Application Layer (IDS)

The application layer can also be referred to as the *presentation layer* and this is where a GUI desktop application written in Java is found. This is the intrusion detection system software interface. This application also allows the user to make various operative selections such as baud rate, port selection as well as to open or close ports in order to control data flow. As long as a gateway is connected and sensors are deployed, the desktop GUI application gives the user power to monitor and view the data that is being sent by the sensors through the gateway. The GUI also issues warning to the user about any activity going wrong with the sensors, as well as all the actions being taken by the internal system.

3) Control Layer

At the control layer, data from sensors is monitored and vital information about the sensors stored in the database. This layer houses the security algorithm, which is the control behind the IDS software. This layer is made up of the algorithm as well as the database. Through this layer, other applications that may need to use the sensor data in a different way from the application layer can connect to the database and gain the information they require. Third party applications also need to connect to and make use of this database, for their data analysis needs as provided by the WSN, as it is will not be easy for two applications to access the same port simultaneously on one computer.

The gateway always connects to any one available port during WSN execution and in the case where a user needs to use another application that reads data as received from the gateway port while simultaneously monitoring the sensors for intrusions, they would need to connect the second application to read data from the data storage of the IDS database such that both applications can benefit. In the control layer, after data has been monitored, and in the case of anomalies being present in the data, the fuzzy logic module detects the anomaly and gives the appropriate response of which the application can communicate to the sensors through OTA.

4) The Application

The application connects to the same port of which the gateway is connected, and receives the frames as they come in and extracts the information needed in order to find anomalies. The anomaly of interest in the case of this research is the ZigBee flood attack. Signs of its possible occurrences are monitored, from the frames being received. The application runs in a loop that checks for this anomaly, for as long as it is in execution.

The solution is composed of three parts; a desktop graphical user Interface, the algorithm and an in-build database. The original algorithm [18] is therefore adapted to include a third phase – resulting in three phases: *learning*, *detection* and *response*.

5) Learning Phase of the Algorithm

The learning phase of the algorithm provides a simple mechanism through which all the nodes connected to the network are discovered and recorded in the database. At the outset, the ZigBee Waspmote gateway [20] sends beaconing signals in order to initiate a way for sensor nodes to discover it. For these sensor nodes to be discoverable, they should be on the same network as the gateway. They should also in turn send their association signals through the same Pan Id and baud rate as expected by the gateway.

The leading task of the algorithm is to discover each and every node that successfully communicates with the gateway and is able to send data to the gateway. These sensor nodes that can connect to the gateway have been given association permission by the gateway to send data and are now part of the network. For as long as the network is up and running, the algorithm goes into an iterative state of learning the current structure of the network with each data transmission instance. The learning phase of the security algorithm therefore mainly records each and every sensor that is making a fresh connection to the network in a database. The nodes are stored using their mac addresses as well as the node ID, both of which form the identification trades for the nodes. The algorithm also takes measures to ensure that there are no nodes that are recorded (in the database) more than once.

ALGORITHM SEGMENT 1: Node DISCOVERY

Let n be the number of nodes that have answered an association call from the gateway
Let i be the node that is being read at a point in time

```
1   For (int i =1; i<n; i++)
        Receive_frame
2   If frame_complete
3       Extract node_mac-address
4       Extract node_id
5   Else discard_frame
6   If mac_address and node_id NOT in Learning_table
    Update learning_table with new mac_address and node_id
            Monitoring and Detection() //calling the detection module
7                                   Else
            discard node_address and node_ID // node is already registered
            Monitoring and Detection()

8                                   End if
9                                   End if
10                                  End while
11          Drop table learning_database
```

The IDS application is meant to be run each time a user wishes to monitor the sensors as deployed in the field. When the application execution is terminated, the learning segment of the IDS algorithm clears the contents of the database, such that it can learn the environment afresh the next time the application executes. This explains the last line (*line 11*) of the algorithm pseudocode, which orders for the contents of the database to be cleared as the usage cycle has timed out. However, if the IDS is run from a server and used for monitoring on a continuous basis, this last line is not to be called upon, until if ever, the IDS is stopped.

6) Detection Phase of the Algorithm

The main purpose of this phase is to detect signs of a flood attack that could be directed at a ZigBee network. Fundamentally, monitoring and detection go together. The following is a segment of the algorithm tasked with monitoring of all of the nodes that are part of the network as well as detecting the nodes that could provide harm to the integrity of the network, and providing corrective solutions for those particular nodes.

TABLE I. ALGORITHM SEGMENT 2: MONITORING AND DETECTION

```
1   Receive_frames
2   Extract node_addresses {Mac_address and node_ID}
3   Compare_address with data in Learning_Database
4   If addresses NOT equal
            Calculate frequency_of_change()
            update table frequency in database}
            If Node_Calculate_frequency ()> 1 then
            Fuzzy_reference () {
                If attack mild

                    issue warning

                Update Database

                Update_environment

                Else if attack_medium {reset node: drop communication}
                Else if attack_severe {isolate node: drop communication}
                } // end of fuzzy module
                Update_environment
5   else
            Update databases  //receive the sensor data and store in database
            Update_environment
6   End if
7   End if
8   End if
```

The algorithm starts by inspecting and comparing the sensor description, with respect to Mac address and node ID, to information that is in the Learning table, in the application database. If one or more of the nodes starts changing their association attributes randomly, the frequency of change is calculated and recorded in a separate table in the database. The fuzzy logic module is also invoked to monitor the level of threat as well as to provide a network correction solution.

TABLE II. ALGORITHM SEGMENT 3: FUZZY LOGIC MODULE

1. ***Define Linguistic variables***
2. *Define membership functions*
3. *Define rules*
4. *Use membership function plots to covert crisp input into fuzzy values*
5. *Evaluate rules in rule base based on fuzzy values*
6. *Summations of results of every rule inferred*
7. ***Deffuzify output (covert output to non-fuzzy values)***

The input for the fuzzy logic in the IDS is the measured frequency of change for any sensor that changes its association attributes. This value is mapped onto fuzzy set membership functions. If a node changes its mac address several times, it actually implies that the node is associating itself afresh to the gateway a lot of times. In this research, this was assumed to be an hypotheses for a possible basis for a flood attack, as one node could request attention from the gateway in as many times as it takes to block the gateway from discovering other legitimate nodes in the time cycle. In the case where there are multiple hops before the message is received at the gateway, it also implies that the sensors relaying the packets to the gateway would be kept awake by one or more rogue sensors and thus filling up their hash tables quicker than they can manage to hold. This is not desirable, as they have limited storage capacity

7) Response Phase of the Algorithm

The response phase works as follows: on the launch of the application, the coordinator sends a beaconing signal of which nodes respond to with a signal showing that they are interested in joining the network. This is entirely the work of the ZigBee protocol. These nodes are individually

programmed such that when association finally happens they send their mac address and node ID address with every packet of message they subsequently sent. These packets therefore contain the information gathered from monitoring a variable of interest as well as the addresses as one message. This data can be anything the sensors are deployed to monitor, be it temperature, soil moisture or motion detected and are not vital to threat detection. The IDS depends on the addresses being packaged together with the sensor data, for monitoring of threats.

As soon as the coordinator has sent an acknowledgement for all the nodes that answered the beaconing request, the security algorithm starts up to intervene and inspect the signals of each and every node in the transmission cycle, before they can be stored and processed. After the process of inspection is complete, the data part of the transmitted signal is stored in a database, where it can be accessed and used for consumption by a third party application that a user may interact with. At this point, rogue nodes would have been found and cleared by the IDS.

III. ALGORITHM ANALYSIS

Levitin [19] defines an algorithm as: "… *a sequence of unambiguous instructions for solving a problem, i.e., for obtaining a required output for any legitimate input in a finite amount of time"*. Performance of algorithms can be based on three metrics: (1) time efficiency – how fast they run; (2) space efficiency – how much memory the need to run; (3) generality – this is in relation to the problem domain and range of inputs it accepts; and (3) a less objective measure of simplicity. In the recapitulation of the analysis framework both time and space efficiencies are the critical measures and are influenced as the functions of the algorithm's input size [19]. Due to the fact that some algorithms' performance differs significantly given same input sizes, it becomes important to determine worst-case, average-case, and best -case efficiencies. Here, the running time (extra memory units consumed) of the algorithm as its input size goes to infinity is measured – this is the principal indicator of the algorithm's efficiency [10] [17] [19]. These measures are represented using the asymptotic notation.

Asymptotic notation was used in this research to analyse the performance of the security algorithm in relation to the input size. This section therefore establishes whether the algorithm will remain robust and fast as input grows. Input in this case refers to the number of nodes that have answered a beaconing request at a particular time interval and are sending data to the coordinator in a transmission cycle. The first task to tackle is finding the number of steps the algorithm takes when trying to establish presence or absence, of security breaches given a reading from the sensor. The learning segment of the algorithm always gets called first when a data reading is sent towards the network and it invokes the detection module and subsequently the fuzzy module, if need be.

When studying the learning segment, the first challenge is the main loop. The rest of the algorithm, (learning and detection) runs from within this loop. When disregarding the entire code, within the loop, it is found that the loop has a total 4 instructions. In the beginning of the algorithm, there is the initialization of i ($i=1$) and the checking, if the loop is still within boundary ($i< n$). In this case i represents the node currently being probed by the algorithm out of a total of all nodes that are part of the transmission cycle. At the end of each run through the body of the loop, there is another check for boundary and the increment of i (i^{++}), which essentially means move to the next node in the sending queue in this current transmission cycle. This gives a total of 4 instructions. When representing this as function, of the number of nodes that can be present at any time in the network, we get $f(n) = 2+ 2n$. That is, 2 instructions that do not directly affect n and two more that do affect n directly.

The next step is to find the Big-OH [10] of the algorithm, which represents the worst case scenario that the system can have. To achieve this, the total number of instructions that can ever happen is counted. This is represented by a case where the instruction steps of the algorithm fire up. These instructions are all the statements in the learning section, detection section as well as the fuzzy algorithm steps. There are, overall, 43 instructions besides those that run the loop. The function at this point can be represented as

$$f(n) = 43 + 2 +2n,$$

which can be simplified to:

$$f(n) =45 +2n.$$

The calculation above shows that there are 45 +2n instructions required when monitoring sensor nodes for security threats in the worst case, where n is the number of nodes sending data towards the gateway at a point in time. However, this function only holds true if we make the following assumptions: first, it is assumed that all instructions such as extraction of node addresses in the learning stage, calculation of frequency of change in detection and monitoring as well as definition of fuzzy linguistics in the fuzzy logic module will happen in constant time for every node. Secondly, it is assumed that all branching statements defined by "if statements" are dealt with in constant time or ***Big-Oh*** (1) for every node being monitored.

The function will therefore hold, with its approximation of 45 instructions when the main loop is not considered. However, all these instructions differ when implemented in different programming languages and are not largely dependent on the growth of the number of nodes that are sending messages to coordinator. In complexity analysis, we consider the number of instructions that grow with the input size. Furthermore, asymptotic notation allows us to filter out slow growing instructions or those instructions that do not directly affect the algorithm as the size of the input grows. This means the 45 instructions remain in constant time as the functions is refined even more. What then remains is

$$f(n)= 2n$$

as it has a directly impact as size grows [10]

Therefore, using the same logic as applied above, the function is reduced to:

$$f(n) = n$$

These properties are not defined in strict mathematics but apply when dealing with asymptotic notation. It is therefore safe to conclude that the behaviour of $f(n)= 45 +2n$ is described by the function $f(n)= n$. Since this was the worst-case scenario, the notation is represented as $O(n)$. The security algorithm has a time complexity of $O(n)$ owing to the fact that at the base of it all it employs a single loop, and

several constant time instructions. An algorithm running in ***O(n)*** is said to be running in linear time, meaning that as the size of the input increases especially for a really large input factor, the running time also increases linearly with the input [10] [17] [19].

IV. Testing and Evaluation

A. IDS Database

A database system is the was used to integrate the major component of the IDS together for the purpose of testing the functioning of the entire system. This database was programmed to not be a standalone database but rather an internal database packaged together with the algorithm, as well as the front end GUI to make one application. It contains a table that was used when the algorithm is learning the state of the network. This table stores records of the nodes that have responded to the beaconing request and have sent association signal to the coordinator to show that they are ready to send in data. Every node gets to have its mac_address and its node address recorded once. If a record of a node_ID is already in the database, it is not recorded again.

Additionally, the database has a table used by the detection section of the algorithm, which is only used to record the nodes that have been changing their mac_address since the first time they were recorded. It counts the number of times a mac_address has changed as "frequency of change". This information is used before the fuzzy logic module can be evoked. This parameter serves as input to the fuzzy logic module. The fuzzy logic module is only called upon by the detection section of the algorithm if this reading has gone above one(1), for any node under inspection. Finally, a third table records the data carried sensors and this table is meant to be accessed by third party applications that need the data for user consumption.

B. Testing

1) Sensor Node Discovery

The first test phase conducted was to check if all the nodes in the network were discoverable and registered in the database, every time the network was re-started. This is important because the anomaly detection algorithm only monitors nodes that are deemed part of the network. That is, only the nodes that have been registered into the database. For this particular assessment, all the nodes were first programmed such that they were sending data at a selected baud rate and transmitting their addresses together with the data they were reporting back to the gateway.

For this test, three (3) Libelium nodes [20] using ZigBee Xbee transmitters were used and a selected number were reprogrammed not comply with the set parameters necessary to be a part of the network with each iteration of the experiment over a period ten (10) times. For the purposes of this experiment, the sensors were measuring temperature and the amount of light. The IDS database was subsequently checked to determine how many nodes had been registered, from those deemed correctly configured.

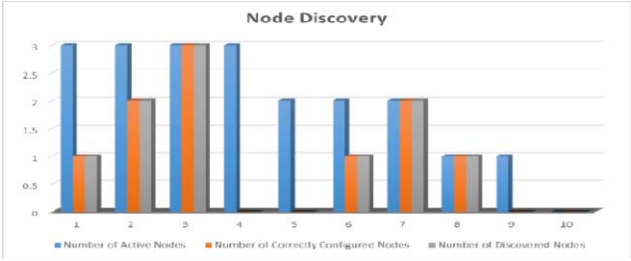


Fig 4: Node Discovery Bar Graph

As indicated in the results in the bar graph in fig.4, the algorithm managed to acknowledge every correctly configured node and recorded it as part of the network. No inactive or correctly configured sensor was every recorded or discovered as part of the network.

2) Anomaly Detection

The second test of the experiment was to test the prowess of the anomaly detection when dealing with simulated ZigBee flood attack. This experiment was conducted by picking out one of the nodes and reprogramming it to behave as though it was trying to cause a flooding of the hash tables in the network. This means that it was reprogrammed to change its mac_address intermittently when sending data to gateway. This was made to be slightly challenging by setting up a random delay period for a node to start behaving maliciously. This was to ensure that a node does not begin to simulate an attack as soon as the node is introduced into the network. The first cycle of transmission for any node is considered as the learning time period and the algorithm is designed not to monitor for denial of service (DOS) flood on the first transmission cycle of a node as no prior information about the node would be in the database.

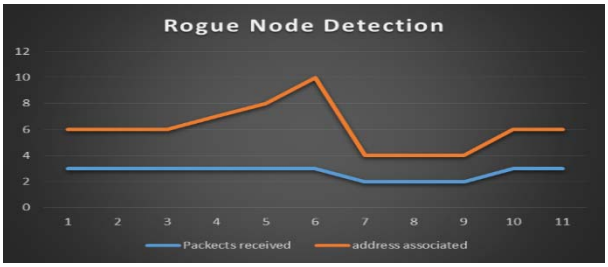


Fig 5: Rogue Node Detections

Fig.5 shows a graph that was used to keep track of the number of packets received in a time period compared to the number of associated addresses behind those packets. The time units between 1 and 3 were a period within which the IDS was learning the network from the moment it was launched. From time sequence 4 to 7, the rogue node started acting up and re-associating itself to the network. This was corrected and the chart shows a normal operational period between time sequences 7 to the end of the simulation period.

Overall, for this experiment, one node was re-programed to be malicious and re introduced to the network. After a period of time it started acting up but was monitored for a limited time period by the IDS until it was reset and dropped from the network. In an effort to find as many flaws as possible, this process was repeated for a minimum of 40 recorded tries, during testing using different sensor nodes as the rogue node for each experiment iteration and at times

varying the number of rogue nodes. Each of those times, the rogue node was sniffed out by the IDS as can be seen in the chart in fig.6.

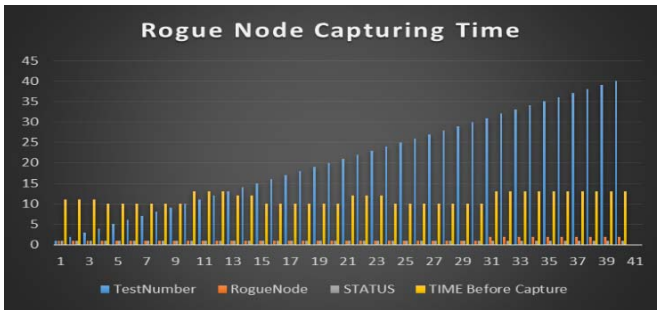


Fig 6: Average Time for Rogue Sensor Detection

V. DISCUSSION AND CONCLUSION

In this paper, an anomaly detection algorithm was adapted, and modified such that it can make use of logic reasoning through fuzzy logic methodologies. This algorithm was then included to be a part of an intrusion detection software and tested for efficiency.

The research focused on the ZigBee protocol and its vulnerability to flood attacks, during node discovery and association to the network. The research then made a novel initiative of taking the same approach that a typical wired network IDS usually takes, of being actually hosted on a server or central computer system that managed such a network. In general, the research introduced an IDS methodology that makes use of a fuzzy logic embedded security algorithm, to monitor for threats in a ZigBee WSN by aggregating, making calculations and assumptions based on the information received from the nodes the moment they send in data from monitoring assignments.

Furthermore, prior to its implementation, the performance of the algorithm was determine using asymptotic analysis where the Big OH was determined. This enabled the determination how well it fares as the number of nodes in a network inevitable explode in size. The asymptotic calculations indicated that the algorithm will execute in linear time as the size of the network grows. This is a good and commendable time complexity for any algorithm, bettered only the time complexity of constant time.

The IDS during simulation was found to be effective in capturing rogue sensor nodes without any problem. The IDS was tested to ascertain its usefulness in a controlled simulation setting. This also gives an idea as to how it will perform when deployed in a real world situation. The algorithm that makes up the IDS executes in an acceptable time complexity meaning it will be very friendly on the RAM of any computer platform it is installed on. The desirable feature of this approach is that there is no extra code work that is left to be done by the nodes except the bits that help them to broadcast their descriptive information with every data packet send, formatted in a way the IDS can make use of.

It is also important to note that a flood attack would not be possible in the first place if this bit of code is excluded by an attacker. This is because of the fact that, any incorrectly configured node is not recognized as part of the network, making this approach again, rather proficient. The IDS was tested against several simulated attacks and it managed to detect and correct them.

For further work, more experiments to determine the performance and accuracy of the algorithm could be carried out on a test bed. It would also add value to this work if the algorithm was incorporated into a real WSN application; for this, plans are underway to incorporate it into a WSN-based weather monitoring system [21] [22].

REFERENCES

[1] A. Gendreau and M. Moorman, "Survey of intrusion detection systems toward an end to end secure internet of things," in *Future Internet of Things and Cloud (FiCloud) 4th International Conference*, Viena, Austria, 2016.

[2] "Towards a Definition of the IoT," May 2015. [Online]. Available: https://iot.ieee.org/definition.html. [Accessed 1 August 2018].

[3] J. Gubbi, R. Buyya, S. Marusic and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future direction," *Future Generation Computer Systems,* vol. 29, no. 7, pp. 1645-1660, 2013.

[4] The Guardian, 2018.[Online]. Available: https://www.theguardian.com/technology/2018/apr/08/facebook-to-contact-the-87-million-users-affected-by-data-breach. [Accessed 31 August 2018].

[5] H. Modares, R. Salleh and Moravejosharieh, "Overview of Security isssues in wireless sensor networks," in *Third International Conference on Computational Intelligence, Modelling & Simulation*, Washington DC, United States of America, 2011.

[6] I. Butun and S. Morgera, "A Survey of Intrusion Detection Systems in Wireless Sensor Networks," *IEEE Communications Survey & Tutorials,* vol. 16, no. 1, pp. 266-282, 2014.

[7] M. Masinde, "Generic Architecture for IoT services;an integrated early drought warning system for sub-Saharan Africa," 2013.

[8] A. Tyagi, J. Kushwah and M. Bhalla, "Threats to security of wireless sensor networks," in *7th International Conference on Data Science & Engineering-Confluence pp. 402-405*, Noita, Uttah Pradesh, India, 2017.

[9] Alrajeh, S. Khan and B. Shams, "Intrusion Detection Systems in wireless sensor networks: a review," *International Journal of Distributed Sensor Networks,* vol. 9, no. 5, 2013.

[10] I. Chivers and J. Sleighthome, "An introduction to Algorithms and the Big O Notation," in *Introduction to Programming with Fortran*, Cham, Springer, 2015.

[11] D. Vishwakarma, "IEEE 802.15.4 and ZigBee: A Conceptual Study," *International Journal of Advanced Research in Computer and Communication Engineering,* vol. 1, no. 7, 2012.

[12] Z. Ang, H. Jin, Z. Fan and Y. Duan, "WSN Node Design and Communication Realization Based on Zigbee Protocol," *Modern Electronics Technique,* vol. 10, pp. 47-50, 2007.

[13] M. Hoque, "An Implementation of Intrusion Detection System Using Genetic Algorithms," *International Journal of Network Security & Its Applications,* vol. 4, no. 2, pp. 109-117, 2012.

[14] Y. Fu, Z. Yan, J. Cao, O. Kone and X. Cao, "An automata based intrusion detection method for internet of things," *Mobile Information Systems,* vol. 2017, no. Article ID 1750637, p. 13 , 2017.

[15] R. Bhaumik, "Mobile Agent based Architecture for Wireless Sensor Networks," in *Seminar on Network Security and Internetworking*, Helsinki, Finland, 2010.

[16] S. Srivastava, N. Gupta, S. Chturvedi and S. Ghosh, "A survey on Mobile Agent based Intrusion Detection Systems," *JCA Proceedings on International Symposium on Devices MEMS, Intelligent Systems & Communication (ISDMISC),* vol. 6, no. 5, pp. 19-24, 2011.

[17] S. Skiena, "Algorithm Analysis," in *The Algorithm Design Manual*, Springer, London, 2012.

[18] B. Stelte and G. Rodosek, "Thwarting Attacks on ZigBee – Removal of the KillerBee Stinger," Universit¨at der Bundeswehr M¨unchen, Neubiberg, 2013.

[19] A. Levitin, Introduction To Design and Analysis of Algorithms, 2/E, India: Pearson Education, 2008.

[20] Libelium.com, Libelium, 21 June 2017. [Online]. Available: http://www.libelium.com/. [Accessed 21 June 2017].

[21] M. Muthoni and B. Antoine, "A Calibration Report for Wireless Sensor-Based Weatherboards," *Journal of Sensor and Actuator Networks,* vol. 4, pp. 30-39, 2015.

[22] M. Masinde, "An innovative drought early warning system for sub-Saharan Africa: Integrating modern and Indigenous approaches," *African Journal of Science, Technology, Innovation and Development,* vol. 7, no. 1, pp. 8-25, 2015.

Board Games in the Computer Science Class to Improve Students' Knowledge of the Python Programming Language

Dawid B. Jordaan
School of Computer Science and Information Systems
North-West University
South Africa
dawid.jordaan@nwu.ac.za

***Abstract*—Increasingly research reveals that well-designed board games inspire learning, encourage communication, teamwork and risk taking. Self-confidence is built as board games empower players. The main objective of this paper is to determine if students will embrace board games as a learning tool to improve their knowledge of the Python programming language. Action research was followed to design and develop a board game for this purpose. The outcomes of this study show that students enjoy the playful learning environment that board games offer and that they welcome board games as a fun and entertaining way to learn. This is a very important result especially for developing countries as board games are inexpensive, interactive and make students mentally and playfully active.**

Keywords—Board games, Computer Science, Python programming

I. Introduction

The information technology industry requires skills from students which include flexibility, critical thinking, problem solving, collaboration and communication. Games can lead to a unique user experience in the learning environment when there is a seamless integration between learning and entertainment. Many elements of game design support a wide range of player abilities, and learning translates to behavior changes. Students learn best when they get the chance to practically experience and integrate knowledge or discuss the knowledge they gain [1]. Properly designed board games are an effective tool for a hands-on and heads-on approach for learning [2]. A board game was designed and developed with the intention to enhance students' Python knowledge and/or Python programming skills. Python is a general-purpose programming language and suitable for introductory programming courses. The main goal of this study is to investigate board games as an interactive learning method to improve students' Python knowledge.

II. Learning Environment

Flexibility, critical thinking, problem solving, collaboration, and communication are skills required by future careers [3]. Graduates need to be prepared to accept positions in the Information Technology (IT) industry that do not exist yet [4]. They need to implement a strategy of lifelong learning as a willingness to continuously develop themselves is a priority to adhere to the fast changing challenges of the IT industry [5-7]. The ways students learn and lecturers teach are influenced by the fast changing information society [3]. Students are familiar with digital tools and learning material which imply a different learning environment and teaching strategy [8, 9]. According to [10] the education systems have primarily remained the same. Resource constrained learning environments are less effective to digitally orientated students [11, 12]. Participatory, sensory-rich and discovery-based learning environments where students are actively involved will appeal to them as technology is part of their lives [13-15]. There is a call for a change in teaching strategy for the new generation of digitally oriented students [16].

Subject delivery can be enhanced by introducing games as a training tool and students can be motivated and perform better when mobile technology is used [17, 18]. A well-designed game allows students to manage and solve problems in an actual situation making students' learning more realistic in terms of the context in which the learning takes place.

III. Computer Programming

The literature reveals a difference between the number of graduates, the number of graduates that are employed and the quantity of unfilled positions in the Information and Communication Technology (ICT) industry [19-21]. In 2006 there were 104 000 vacancies in the ICT field in South Africa and in 2007 in the region of 200 000 unemployed ICT graduates, with an assortment of qualifications [22, 23]. The ongoing enrolment crisis in computer science can be blamed on unrelated course material and learning content. This is a serious concern as we live in a technology driven society. A study by [24] reveals that students loose interest and a negative mindset is developed for technical subjects such as programming when they experience the subject as difficult to learn. This creates a perception that programming, which is an integral part of almost all computer science curricula, is a difficult skill to master. This is reported as one of the reasons for the decline in interest in studying computer science [24, 25]. The ability to solve problem in the programming environment requires higher order thinking skills which is expected from students in today's digital environment [26-28]. Students are more determined to solve problems when they are motivated and interactively involved and when they can relate to the context in which programming problems are presented [27, 29, 30].

IV. Technology in the Computer Environment

A learning environment using technology can benefit students as they are better motivated by technology [31, 32]. Technology fast-tracks and improves basic skills, is extremely motivational since it offers comfort to digitally oriented students, facilitates new fields through games, simulations, three-dimensional models, etc., and prepares students for the work place from a young age [33]. Portability of mobile

technology allows learning anywhere, anytime and at the student's own pace [34].

A. Games

Many researchers agree on the idea of motivational purposes of a playful approach in education [35]. The contribution to an effective and powerful learning environment of digital games leads to the integration of digital games into the learning environment [36, 37]. The popularity of video games have increased and the extraordinary motivation and user participation that games offer have motivated researchers to consider the serious games concept in hopes of enticing learners [38, 39]. Serious games offer a playful element complimented by aspects like interactivity, rewards, challenges and scores [40, 41]. These attributes together with the playful element of serious games may yield positive results in terms of learning and motivate students as they easily accept new technology [42]. The use of serious games in the learning environment lead to a unique user experience when there is a seamless integration between learning and entertainment [35].

Many definitions of serious games are reported in the literature. The majority of these definitions concur that the primary objective of a serious game is not just entertainment [43]. The fun part is important, but the focus of serious games should be on problem-solving and elements of learning [44].

B. Board Games

The Merriam Webster dictionary describe a board game as "a game of strategy (such as checkers, chess, or backgammon) played by moving pieces on a board [45] describes." The Collins dictionary explains a board game as "a game such as chess or backgammon, which people play by moving small objects on a board [46]." Board games date back to 5000 BC [47]. Normally a board game is played on a flat surface like a table and the board is the center point. Different approaches are used to achieve the goal of a board game, which includes pure strategy, an element of chance or no element of skill but purely chance. The educational property of board games may provide players with some type of knowledge through game play [48].

V. Study Objective

Studies that focus on games in education are aimed at younger age groups with largely positive attitudes towards games in class [49]. It is not well documented whether the idea of serious games or board games in the computer science class will be embraced by students in tertiary education as they may have different perceptions, prospects, requirements and profiles. The following objectives and research questions have been set for this study.

A. Primary Objective

The primary objective of this study is to investigate the use of board games in the computer science class to improve students' knowledge of the Python programming language.

B. Research Quiestions

To achieve the primary objective the research questions were formulated as follows.

Research question 1. Will students embrace board games in the computer science class to improve their knowledge of the Python programming language?

Research question 2. Will lecturers use board games in the computer science class as a tool to improve students' knowledge of the Python programming language?

C. Secondary Objective

To achieve the primary objective the secondary objective was set to find board games that focus on improving students' knowledge of the Python programming language and use them to answer the research questions. The secondary objective was set to design and develop a board game suitable for the task because searches did not reveal any existing board game appropriate to answer the research questions for this study.

VI. Methodology

As no board game suitable for this study could be found the solution was to design and develop a board game to assist in answering the research objective. For this action research was used because action research is especially suitable in settings where educators and practitioners want to reflect on their own practices [50]. Action research is practical, participatory, starts with a problem and attempts to find a solution to the problem. Action research attempts to improve professional practice through a cycle of reflective stages to develop expertise; to understand the practice and the development of a well specified theory of action; and through participatory research improvement of the community [51].

Action research is a cyclic process executed in five phases [52]. At first the research must be placed in an environment where the specification and agreement are defined, which determine the research environment. When the research environment is determined the following five distinct phases are repeatedly executed in this environment: diagnosing, action planning, action taking, evaluating and specifying learning.

Diagnosing. In this phase the primary reasons why the organisation wants to change are identified.

Action planning. This phase requires a partnership between researchers and practitioners. Together they determine actions to solve or at least lessen the primary problems identified during the diagnosing phase.

Action taking. The identified actions are implemented during the action taking phase.

Evaluating. The outcomes of the completed actions are evaluated during this phase. The evaluation determines if the theoretical implementation of the actions was comprehended and whether these effects solved the problems. If the desired changes were not successful a new framework should be created to correct the situation in the next iteration of the action research cycle.

Specifying learning. The specified learning phase is formally executed at the end and is usually an on-going process. Action research provides new knowledge that benefit three audiences, namely: organisational norms are restructured; unsuccessful changes may provide a basis for additional action; and the success or failure of the theoretical

framework compiled by this research benefits future research [53].

Following these five phases the board game was developed in two action research cycles.

VII. The Board Game

Fig 1 depicts the board game designed, developed and used in this study. The game is suitable for two to ten players but preferably four to six payers should play. The game consists of a board with a circle drawn on it as shown in Figure 1. The circle is divided in nine sections in the form of a pie chart. On each section are instructions to be executed. A spinner is mounted on the board and when turned will stop on a section of the circle. Accompanying the board is a tablet. An app on the tablet host a database with approximately 400 Python related questions. The questions are color-coded as orange, blue, green and red – ranging from easy (orange), moderate (blue), not so easy (green) to challenging (red). Depending on the instruction on the circle segment a color is selected, the app will randomly choose a Python question from that specific colour category, which must be answered in the allotted time for that specific question displayed by the app. On request the app displays a Hint, the Answer and a Discussion of the answer. Players need to answer the questions and are rewarded with bit_dollar$.

Playing the game. The first player (player1) spins the spinner which will stop on a segment of the board. Depending on the segment where the spinner stopped player1 reads the instruction on the segment to the other players. Instructions are one of the following: select a blue question; select a red question; select a green question; select an orange question; select any question (if one player answers the question wrong nobody gets a bit_dollar$–that is, nobody gets rewarded); grab one bit_dollar$ from any player and spin again; donate one bit_dollar$ to any player and spin again. Player1 taps the colour-coded question on the tablet and a question is revealed. All the other players now individually answer the question using a pen and notepad. When the time indicated on the tablet has elapsed, player1 collects all the answers and marks them using the Answer button on the tablet. Players get rewarded with bit_dollar$ for correct answers. If anybody answer the question incorrectly the group must discuss the question as well as the answer using the Discussion button on the tablet. When everybody is satisfied the game continues with the next player to spin the spinner. The players decide when to stop the game and the player with the most bit_dollar$ wins.

VIII. Data Collection

Interviews, a purposeful discussion between two or more people, can help gather data that are valid, reliable and relevant to research questions and objectives [54]. Structured, semi-structured and unstructured interviews are different forms of interviews available to the researcher. During structured interviews the researcher use the same questionnaire for all participants. During semi-structured interviews the researcher has a list of themes and questions that need to be discussed, but the flow and order in which the participants respond to these themes or questions can vary from participant to participant. Unstructured interviews are informal and used to explore a general area in which researchers are interested in, in depth.

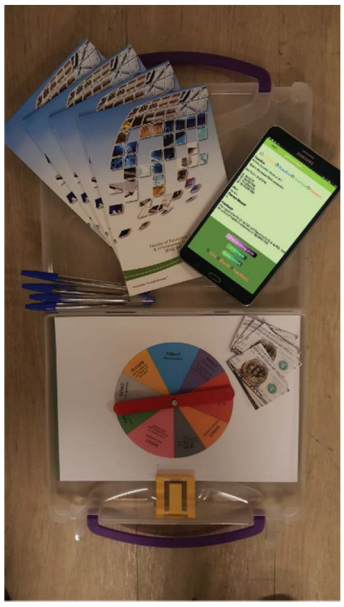


Figure 1. The board game developed for this study (Source: Author)

To test the board game and the game mechanics two non-academic people played the game. A few suggestions were made, and the board game was finalized. Lecturers and students from a Higher Education Institution participated in this study. All lecturers have experience in teaching programming at first year level. Students were enrolled for the Python programming class and were first-year students as Python is offered at first year level and the idea was to let them experience the board game while they study the programming course. Eight lecturers and eight students volunteered to play the board game. Participants were divided into four groups (two lecturer groups and two student groups) with four players in each group. The four groups played the game and afterwards semi-structured interviews were conducted with each group. Groups were asked to comment on issues like enjoyment, collaboration, communication, socialisation, involvement, participation and subject content. Players could raise their opinions freely and openly and comments were written down on paper and at the end of the discussions summarised to the agreement of the group members.

IX. Findings

From the semi-structured interviews the following findings were formulated (see Appendix for semi-structured questions).

- *Enjoyment*. All groups highlighted the fun and enjoyment they had playing the game and they enthusiastically engage themselves without any hesitation. Players will enjoy learning if they are actively involved as stated by [13].
- *Communication*. Players agree that the option to discuss the answers to the questions could improve knowledge retention which is in line with the findings of [1]. All groups mentioned that communication skills in general could also improve.
- *Socialisation*. Players agree that by playing the board game group engagement will enhance socialisation. Playful socialisation and spontaneous engagement are also positive points raised by the groups.

- *Subject content*. The board game addresses only one specific subject which force players to focus on content. Players agreed that, in a playful way, the board game forced them to be engaged with subject content. Players mentioned that by discussing the answers using the feedback option of the app, players get the opportunity to learn to communicate subject-related terms.
- *Collaboration*. Players were unanimous that this version of the board game does not address collaboration effectively.
- *Hands-on heads-on*. There was agreement that although the questions were appropriate and not too difficult, they had to concentrate to get the correct answer in the designated time–*heads-on* and *hands-on* is achieved as the game involves all players. This is in line with researchers who states that board games are effective tools for *hands-on heads-on* activities [2].
- *Interactive involvement*. All groups agree that the tangible elements of the game ensure interactive involvement. This corresponds with research stating that learning environments where students are actively involved will appeal to them [13-15].
- *Motivation*. Motivational purpose of a playful approach as stated by [35] was confirmed by the majority players (students and lecturers) as they mentioned the possibility to use and play board games in class.

X. Conclusion

Well-designed board games are effective and valuable learning tools and adults, like children, enjoy learning when a fun component is attached to it [2]. This joyful experience can be extended with team-based board games designed on specific topics to provide information, promote teamwork and debates, present challenges and provide entertainment and engagement. This is likely because proper design integrates all elements found to encourage effective learning. Overall, students and lecturers are positive that this board game may improve their knowledge of the Python programming language in a playful manner and that it is an inexpensive, interactive learning method to stimulate heads-on and hands-on learning activities [2]. This observation answers both the research questions "Will students embrace board games in the computer science class to improve their knowledge of the Python programming language?" and "Will lecturers use board games in the computer science class as a tool to improve students' knowledge of the Python programming language?"

XI. Recommendations

It is recommended that the use of board games should be seriously considered in the Computer Science class.

References

[1] S. Ozsoy, G. Ozsoy, and H.G. Kuruyer, *Turkish pre-service primary school teachers' environmental attitudes: Effects of gender and grade level*. 7th International Conference on Education, Asia-Pacific Forum on Science Learning & Teaching. Place of publication: Samos,

[2] N. Treher, "Learning with Board Games," *The Learning Key, Inc.* [Online]. Available: https://www.thelearningkey.com/.../Board_Games_TLKWhitePaper_May16_2011.pdf. [Accessed: 25 November 2017].

[3] K. Gravemeijer, "Aiming for the 21st century skills," *International Journal for Mathematics in Education,* vol. 4, pp. 30-42, 2012

[4] I. Wolfe, "65 Percent of today's students will be employed in jobs that don't exist yet." 2013 [Online]. Available: https://www.successperformancesolutions.com/65-percent-of-todays-students-will-be-employed-in-jobs-that-dont-exist-yet/. [Accessed: 24 May 2017].

[5] C.M. Kaiser, and M.A. Wisniewski, "Enhancing student learning and engagement using student response systems," *Social Studies Research and Practice,* vol. 7, (2), pp. 137-149, 2012

[6] G. Rayner, and T. Papakonstantinou, "Employer perspectives of the current and future value of STEM graduate skills and attributes: An Australian study," *Journal of Teaching and Learning for Graduate Employability,* vol. 6, (1), pp. 100-115, 2015

[7] Right Management,"The Flux Report: Building a resilient workforce in the face of flux," [Online]. Available: http://www.rightmanagement.co.uk/wps/wcm/connect/350a18c6-6b19-470d-adba-88c9e0394d0b/Right+Management+Flux+Report+Spread.pdf?MOD=AJPERES. [Accessed: 25 October 2017].

[8] J. Bidarra, M. Figueiredo, and C. Natálio, "Interactive design and gamification of eBooks for mobile and contextual learning," *International Journal of Interactive Mobile Technologies (iJIM),* vol. 9, (3), pp. 24-32, 2015

[9] M. Prensky, "The games generations: How learners have changed," in Digital game-based learning, McGraw-Hill, 2001, pp. 1-26

[10] C. Conneely, C. Girvan, and B. Tangney, "An exploration into the adaption of the Bridge21 model for 21st century learning in Irish classrooms," Dublin, T.C., Dublin, 2012.

[11] M. Cunningham, *Factors impacting on adoption of Technology-enhanced Learning techniques by universities in Nairobi, Kenya*. 2015 IEEE International Symposium on Technology in Society (ISTAS) Place of publication: Dublin, Ireland, 2015.

[12] P. Rooney, "Theoretical Framework for Serious Game Design: Exploring Pedagogy, Play and Fidelity and their Implications for the Design Process," *Journal of Game-Based Learning,* vol. 2, (4), pp. 41-60, 2012

[13] L.A. Annetta, J. Minogue, S.Y. Holmes, and M.T. Cheng, "Investigating the impact of video games on high school students' engagement and learning about genetics," *Computers & Education,* vol. 53, (1), pp. 74-85, 2009

[14] D. Heersink, and B.M. Moskal, "Measuring high school students' attitudes toward computing," *SIGCSE,* vol. 10, pp. 10-13, 2010

[15] L. Husain, "Getting serious about math: serious game design framework & an example of educational game," Lund University Libraries, 2011.

[16] M.J. Zeeman, "Modelling the factors that influence computer science students' attitude towards serious games in class," Ph.D. dissertation, North-West University, Vanderbijlpark, South Africa, 2014.

[17] S.A.M. Yusof, S.H.M. Radzi, S.N.S. Din, and N. Khalid, *A study on the effectiveness of task manager board game as a training tool in managing project*. AIP. Place of publication: Kedah, Malaysia, 11–13 April 2016.

[18] A. Heck, P. Boon, C. Bokhove, and G. Koolstra, "Applets for learning school algebra and calculus: Experiences from secondary school practice with an integrated learning environment for mathematics," [Online]. Available: https://eprints.soton.ac.uk/id/eprint/348449, [Accessed: 15 June 2017].

[19] K. Gareis, T. Hüsing, B. Strahil, I. Bludova, C. Schulz, and W.B. Korte, "E-Skills for jobs in Europe: Measuring progress and moving ahead," Empirica, Bonn, Germany, 2014.

[20] S. McLaughlin, M. Sherry, E. Doherty, M. Carcary, C. Thornley, Y. Wang, M. Kolding, G. Cattaneo, W.B. Korte, S. Birov, T. Heusing, M. Steinback, F. Fanning, E. Valdelin, M. Stork, and M. Krmek,"e-skills: The international dimension and the impact of the globalisation," European Commission. [Online]. Available: http://www.cepis.org/media/FINAL_INTERNATIONAL_e_Skills_report_Aug_141.pdf. [Accessed: 26 October 2017].

[21] A. Schofield,"2014 JCSE ICT skills survey report," Joburg Centre for Software Engineering. [Online]. Available: http://windowsitpro.com/industry/2014-it-skills-and-salary-report. [Accessed: 25 October 2017].

[22] C. Pop, and N. Barkhuizen, "The relationship between skills training and retention of graduate interns in a South African information, communication and technology company," *Literacy Information and Computer Education Journal,* vol. 1, (2), pp. 113-122, 2010

[23] JIPSA,"Towards a JIPSA ICT skills intervention strategy," [Online]. Available: http://www.cs.ru.ac.za/ictskills/DilloLehlokoePresentation.pdf. [Accessed: 12 November 2017].

[24] F.L. Khaleel, N.S. Ashaari, T.S. Meriam, T. Wook, and A. Ismail, *The study of gamification application architecture for programming language course*. Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication. Place of publication: Bali, Indonesia, 2015.

[25] A. Peters, and A. Pears, *Students' experiences and attitudes towards learning computer science*. Proceedings of the Frontiers in Education Conference (FIE). Place of publication: Seattle, WA, USA,

[26] S. Kurkovsky, "Mobile game development: improving student engagement and motivation in introductory computing courses," *Computer Science Education,* vol. 23, (2), pp. 138-157, 2013

[27] B. Tan, and H. Rahaman, *Virtual heritage: Reality and criticism*. Proceedings of the 13th International CAAD Futures Conference. Place of publication, 2009.

[28] S. Nag, J.G. Katz, and A. Saenz-Otero, "Collaborative gaming and competition for CS-STEM education using SPHERES Zero Robotics," *Acta Astronautica,* vol. 83, pp. 145-174, 2013

[29] S. Shabanah, and J.X. Chen, *Simplifying algorithm learning using serious games*. Proceedings of the 14th Western Canadian Conference on Computing Education. Place of publication: Burnaby, British Columbia, Canada, 2009.

[30] I.A. Stanescu, A. Stefan, and I. Roceanu, "Interoperability in Serious Games," *Conference proceedings of eLearning and Software for Education (eLSE),* vol. 1, (1), pp. 19-24, 2011

[31] G. Schwabe, and C. Göth, "Mobile learning with a mobile game: design and motivational effects," *Journal of computer assisted learning,* vol. 21, (3), pp. 204-216, 2005

[32] C. Mbogo, E. Blake, and H. Suleman, "Initial Evaluation of a Mobile Scaffolding Application That Seeks to Support Novice Learners of Programming." in *Paper presented at the International Conference on Mobile Learning.* Place of publication: Madrid, Spain, 2014, pp. 8.

[33] E.C. Coughlin, and C. Lemke, "Professional competency continuum: Professional skills for the digital age classroom," vol. pp.1999

[34] A. Bartel, and G. Hagel, *Engaging students with a mobile game-based learning system in university education*. Paper presented at the Global Engineering Education Conference (EDUCON). Place of publication, 3-5 April 2014.

[35] P.T. De Villiers, and A.S. Blignaut, "Design evaluation of a forklift serious game," [Online]. Available: http://dergipark.gov.tr/download/article-file/257166, [Accessed: 5 December 2017].

[36] S. Heintz, and E.L.C. Law, "Evaluating design elements for digital educational games on programming: A pilot study." in *26th Annual BCS Interaction Specialist Group Conference on People and Computers.* Place of publication: Birningham, UK, 2012, pp. 245-250.

[37] M. Wrzesien, and M. Alcañiz Raya, "Learning in serious virtual worlds: Evaluation of learning effectiveness and appeal to students in the e-junior project," *Computers & Education,* vol. 55, (1), pp. 178-187, 2010

[38] M. Zyda, "From visual simulation to virtual reality to games," *Computer,* vol. 38, (9), pp. 25-32, 2005

[39] C.H. Su, and C.H. Cheng, "A mobile gamification learning system for improving the learning motivation and achievements," *Journal of Computer Assisted Learning,* vol. 31, (3), pp. 268-286, 2015

[40] J.J. Vogel, A. Greenwood-Ericksen, J. Cannon-Bowers, and C.A. Bowers, "Using virtual reality with and without gaming attributes for academic achievement," *Journal of Research on Technology in Education,* vol. 39, (1), pp. 105-118, 2006

[41] A.L. Krassmann, L.N. Paschoal, A. Falcade, and R.D. Medina, *Evaluation of Game-Based Learning Approaches through Digital Serious Games in Computer Science Higher Education: A Systematic Mapping*. Proceedings of the 14th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames). Place of publication: Piaui, Brazil, 2015.

[42] T. Wang, and Y. Tseng, "An empirical study: Develop and evaluation a mobile serious game on environmental education," *In Computer Science & Education (ICCSE), 2014 9th International Conference on,* vol. pp. 4, 2014

[43] D. Djaouti, J. Alvarez, and J.P. Jessel, "Classifying serious games: the G/P/S model," *Handbook of research on improving learning and motivation through educational games: Multidisciplinary approaches,* vol. 2, pp. 118-136, 2011

[44] T. Susi, M. Johannesson, and P. Backlund, "Serious games: An overview," Skövde, U.o., Sweden, 2007.

[45] Merriam-Webster, "Board game," Merriam Webster Dictionary, [Online], Available: https://www.merriam-webster.com/dictionary/board%20game, Accessed: 25 January 2018.

[46] Collins-Dictionary, "Board games," Collins English Dictionary, [Online], Available: https://www.collinsdictionary.com/, Accessed: 25 January 2018.

[47] P. Attia, "The Full History of Board Games," *The Startup*. 2016 [Online]. Available: https://medium.com/swlh/the-full-history-of-board-games-5e622811ce89 [Accessed: 28 May 2017].

[48] S. Joseph, and L. Diack, "Playing interprofessional games: reflections on using the Interprofessional Education Game (iPEG)," *Journal of Interprofessional Care, Informa UK Ltd,* vol. 29, (3), pp. 2, 2014

[49] M. Papastergiou, "Digital game-based learning in high school computer science education: Impact on educational effectiveness and student motivation," *Computers & Education,* vol. 52, (1), pp. 1-12, 2009

[50] M.E. George, *Action Research: A guide for the Teacher Researcher*. Upper Saddle River, New Jersey, Prentice Hall, 2003, pp. 235.

[51] J.W. Creswell, and V.L.P. Clark, *Designing and conducting mixed methods research*. India, Sage Publications, Inc, 2007, pp. 520.

[52] G. Susman, and R. Evered, "An Assessment of The Scientific Merits of Action Research," *Administrative Science Quarterly,* vol. 23, (4), pp. 582-603, 1987

[53] R.L. Baskerville, "Investigating information systems with action research," *Communication of the association for information systems,* vol. 2, (19), pp. 2-31, 1999

[54] M. Saunders, P. Lewis, and A. Thornhill, *Research methods for business students*. England, Pearson Edcuation, 2009, pp. 614.

APPENDIX

As previously mentioned semi-structured interviews are interviews where the researcher has a list of themes and/or questions that need to covered. During the semi-structured interviews respondents were asked to comment on the folowing themes as they experienced it while playing the game; enjoyment, communication, socialisation, subject content, collaboration, hands-on heads-on experience, interactive involvement and motivation. Lecturers were also asked if they will consider it to use the game in class.

Predator-Prey Model for Illicit Drug Consumption

1st Jean-Marc Ginoux
LIS UMR CNRS 7020
CS 60584, 83041,
Toulon Cedex 9, France
ginoux@univ-tln.fr

2nd Roomila Naeck
LIS UMR CNRS 7020
CS 60584, 83041,
Toulon Cedex 9, France
roomila.naeck@gmail.com

3rd Yusra Bibi Ruhomally
Department of Mathematics
University of Mauritius
Réduit, Mauritius,
bibi.ruhomally2@umail.uom.ac.mu

4th Muhammad Zaid Dauhoo
Department of Mathematics
University of Mauritius
Réduit, Mauritius,
m.dauhoo@uom.ac.mu

***Abstract*—Mathematical modeling of illicit drug consumption is a very difficult and complex problem. To this aim predator-prey models have been used at the end of the nineties. Then, in 2013 a model called NERA has been built to describe the dynamics of the non (N), experimental (E), recreational (R) and addict (A) user categories, respectively, within a given population. However, the original NERA model didn't involve limitation in drug consumption and was consequently unable to transcribe the periodic evolution of each category. So, we have modified this model by analogy with the classical predator-prey models and while considering non-users (N) as prey and users (E, R and A) as predator. Then, using data from the state of Colorado and Washington and a genetic algorithm, we calibrated our predator-prey NERA models by estimating their parameters sets. This allowed us to account for the periodic evolution of each category. Then, by considering that the population of Nonusers increases in proportion of the demography we highlighted chaotic regions within which the evolution of the variables becomes unpredictable. Thus, it appears that our validated NERA model can be a precious tool in forecasting of illicit drug consumption and can be of substantial interest to policy-makers in the problematic of illicit drug consumption.**

***Index Terms*—UJT, memristor, torus breakdown, bifurcation**

I. Introduction

Since the beginning of the nineties, many continuous time models have been proposed to describe the dynamics of drug consumption [1]–[5], [12], [14]. They mainly consisted in second order nonlinear models involving two variables which were used for predicting the long term dynamics of addicts and dealers of a large drug market, like that of an entire country. Few years later, Gragnani *et al.* [13] proposed an extension to such models by adding a third ordinary differential equation to the original two-dimensional dynamical systems. This third variable represented the enforcement exerted by the authorities. Let's notice on the one hand that these models already used limitations in the growth and decay of each variable (at least for the two first) and on the other hand that, Gragnani *et al.* [13] proved the existence of slow-fast limit cycles according to the singular perturbation theory as solution of their three-dimensional dynamical system. Recently, Dauhoo *et al.* [6] have considered that drug users are generally classified into three main categories, depending on their consumption frequency and the control they have over the drug, i.e., Experimental (E), Recreational (R) and Addicts (A) users. By adding a fourth variable, i.e., the Non (N) users to these first three ones, they proposed the NERA model. Although this four-dimensional dynamical system takes into account the mutual influence that drug users (E, R and A) can have on non-users (N) and on each other it does not contain any limitation in the growth and decay of each variable. Thus, no oscillatory or chaotic regime could be observed. That's the reason why we have proposed to modify Dauhoo's NERA model [6] by introducing a limitation in each "functional response". In their paper, Dauhoo *et al.* [6] wrote the following sentence: "Anyone could be a 'prey' to illicit drugs". This led us to make the analogy with predator-prey models.

II. General Predator-Prey NERA Model

This deterministic model aims at transcribing into mathematical functions variations of the number of individuals of each group due to their interactions with others. We make the assumption that such interactions are mainly characterized by the influence that individuals of one group may exert on the others. Thus, we consider that people influenced by a group leaves the group to which they belong to join the group which has influenced them. As a consequence, some individuals disappear of one group and appear in another. So, by analogy with the predator-prey models used for a long time in Theoretical Ecology [22], [27], these influences, which cause such appearances and disappearances, *i.e.*, increases and decreases (variations) of the number of individuals of each group can be regarded as predation of on group on another. In the predator-prey model that we propose, Non-users (N) represent prey for all other groups (E, R and A). Then, Experimental-users (E) are predator of both Recreational-users (R) and Non-users (N) while Addict-users (A) are predator of all R, E and N-users. According to Kuznetsov [20], for the model to be realistic it is necessary to introduce a "stabilizing effect of the competition among prey and a destabilizing effect of the predator saturation". To this aim, we have used two different types of "functional responses" for the growth of prey (N) and for the growth of the predators (E, R, A). We have first considered that, in the absence of any predator, prey growth (N) must be limited. Such limitation or stabilizing effect is generally introduced by using the *logistic law* introduced by Verhulst [30]. Concerning the predators (E, R, A), the saturation of the predator rate (destabilizing effect) can be modeled with the classical Holling type II "functional response" [16], [17]. Of course, various other "functional responses" could have been chosen to this aim [11]. Let's notice that such

saturation in the predator rate represents a limitation in the influence of each predator group on the others. At last, we consider that in the absence of its predators, the number of individuals of one group (E, R, and A except N) can decrease by a kind of "natural mortality" which can correspond to individuals leaving this group. In the most dramatic case of drug addict, this could be due to overdose. So, in this work we propose to used both *logistic law* and Holling type II functional response for modeling the influence that exert the predators A, R and E on each others and on the prey N.

A. Model equations

So, we have the following system of ordinary differential equations:

$$\begin{aligned}
\frac{dN}{dt} &= \beta_1 N(1-N) - r_1\frac{NE}{h+N} - \alpha_1\frac{NA}{h+N} - \alpha_2\frac{NR}{h+N},\\
\frac{dE}{dt} &= r_1\frac{NE}{h+N} - r_2\frac{ER}{h+E} - \beta_2 E - \gamma_1\frac{EA}{h+E},\\
\frac{dR}{dt} &= r_2\frac{ER}{h+E} - \beta_3 R - r_3\frac{RA}{h+R} + \alpha_2\frac{NR}{h+N},\\
\frac{dA}{dt} &= r_3\frac{RA}{h+R} - \beta_4 A + \alpha_1\frac{NA}{h+N} + \gamma_1\frac{EA}{h+E},
\end{aligned} \tag{1}$$

where β_1 is the growth rate of the population of the prey (N) in the absence of any predator (E, R, A), β_i with $i = 2, 3, 4$ are the "natural mortality" of each predator (E, R, A) in the absence of all others and r_i with $i = 1, 2, 3$ is the predation rate of A on R, R on E and E on N respectively. α_1 and γ_1 represent the predation rate of A on N and E respectively while α_2 is that of R on N. Thus, in this four-dimensional dynamical system, a set of 11 positive parameters: $(\beta_1, \beta_2, \beta_3, \beta_4, r_1, r_2, r_3, \alpha_1, \alpha_2, \gamma_1, h)$ is used.

Remark. Let's notice that for each Holling type II functional response a different half saturation h could have been chosen. Nevertheless, the aim of this work is to propose the most simple and consistent model for illicit drug consumption.

The sociological meaning of each parameter used in this model (1) is given in Table II.

TABLE I: Interpretation of the parameters in *NERA* model

Parameter	Sociological Meaning
r_1	Influence rate of $E(t)$ on $N(t)$
r_2	Influence rate of $R(t)$ on $E(t)$
r_3	Rate at which recreational users change to addicts
α_1	Influence rate of $A(t)$ on $N(t)$
α_2	Influence rate of $R(t)$ on $N(t)$
γ_1	Influence rate of $A(t)$ on $E(t)$
β_1	Rate of moving in and out of the Nonuser category
β_2	Rate at which experimental users quit drugs
β_3	Rate at which recreational users quit drugs
β_4	Rate at which addicts quit drugs

B. Dynamic aspects

Due to the presence of the Holling type II functional responses, the determination of the fixed points of this four-dimensional dynamical system (1) is not trivial while using the classical nullclines method. Nevertheless, by posing $E = R = A = 0$ two obvious fixed points can be easily found:

$$O(0,0,0,0) \quad \text{and} \quad I_1(1,0,0,0).$$

By posing $R = A = 0$, a third fixed point can be also obtained:

$$I_2(N_2^*, E_2^*, 0, 0) = \left(\frac{\beta_2 h}{r_1 - \beta_2}, \frac{r_1 - \beta_2(1+h)}{(r_1-\beta_2)^2}\beta_1 h, 0, 0\right)$$

It follows that this fixed point I_2 is positive provided that:

$$r_1 - \beta_2 > 0 \text{ and } r_1 - \beta_2(1+h) > 0. \tag{2}$$

Then, by posing $E = R = 0$, a fourth fixed point can be also obtained:

$$J_2(N_2^*, 0, 0, A_2^*) = \left(\frac{\beta_4 h}{\alpha_1 - \beta_4}, 0, 0, \frac{\alpha_1 - \beta_4(1+h)}{(\alpha_1+\beta_4)^2}\beta_1 h\right)$$

It follows that this fixed point J_2 is positive provided that:

$$\alpha_1 - \beta_4 > 0 \text{ and } \alpha_1 - \beta_4(1+h) > 0. \tag{3}$$

Finally, while posing $A = 0$, a fifth fixed point $I_3(N_3^*, E_3^*, R_3^*, 0)$ can be determined but its expression is too long to be written here.

Remark. Let's notice that all fixed points depend on parameters (β_1).

Following the works of Freedman [8], the system (1) may be written as:

$$\begin{aligned}
\frac{dN}{dt} &= Ng(N) - (r_1 E + \alpha_1 A + \alpha_2 R)\, p_1(N),\\
\frac{dE}{dt} &= E(-\beta_2 + r_1 p_1(N)) - (r_2 R + \gamma_1 A)\, p_2(E),\\
\frac{dR}{dt} &= R(-\beta_3 + r_2 p_2(E) + \alpha_2 p_1(N)) - r_3 A p_3(E),\\
\frac{dA}{dt} &= A[-\beta_4 + r_3 p_3(R) + \alpha_1 p_1(N) + \gamma_1 p_2(E)],
\end{aligned} \tag{4}$$

where $g(N) = \beta_1(1-N)$, $p_1(N) = \frac{N}{h+N}$, $p_2(E) = \frac{E}{h+E}$ and $p_3(R) = \frac{R}{h+R}$. Such a formulation will simplify the computation of the eigenvalues of the functional Jacobian matrix .

1) Eigenvalues at O(0,0,0,0): In this case, the functional Jacobian matrix is diagonal. Thus, the four eigenvalues are:

$$\lambda_1 = \beta_1 \ ; \ \lambda_2 = -\beta_2 \ ; \ \lambda_3 = -\beta_3 \ ; \ \lambda_4 = -\beta_4.$$

It follows that the origin O is a *saddle*.

2) Eigenvalues at I_1(1,0,0,0): In this case, the functional Jacobian matrix is also diagonal. Thus, the four eigenvalues are:

$$\lambda_1 = -\beta_1 \; ; \; \lambda_2 = -\beta_2 + \frac{r_1}{h+1} \; ;$$
$$\lambda_3 = -\beta_3 + \frac{\alpha_2}{h+1} \; ; \; \lambda_4 = -\beta_4 + \frac{\alpha_1}{h+1}.$$

According to conditions (2-3), both λ_2 and λ_3 are positive. So, whatever the values of the parameters, it follows that I_1 is also a *saddle*.

3) Eigenvalues at I_2(N_2^*, E_2^*, 0, 0) and at J_2(N_2^*, 0, 0, A_2^*): Although the four eigenvalues evaluated at I_2 and J_2 are too long to be expressed, two of them contain a square root. So, according to the choice of the parameters, these eigenvalues may be complex conjugate. Thus, if we assume that the expression in the square root is negative, we can look for the sign of the remaining part which can be considered as the *real part* of these eigenvalues. Such a real part is positive provided that:

$$0 < r_1 - \beta_2 (1+h) - h r_1 \quad \text{and} \quad 0 < r_1 - \beta_2,$$
$$0 < \alpha_1 - \beta_4 (1+h) - h\alpha_1 \quad \text{and} \quad 0 < \alpha_1 - \beta_4.$$

Combining these conditions with the previous one (2-3), we find that

$$\begin{aligned} 0 < h r_1 < r_1 - \beta_2 (1+h), \\ 0 < h\alpha_1 < \alpha_1 - \beta_4 (1+h). \end{aligned} \tag{5}$$

It follows that I_2 and J_2 are a *saddle-foci* (two eigenvalues are complex conjugate with positive real parts and the two others eigenvalues are real).

4) Eigenvalues at I_3(N_3^*, E_3^*, R_3^*, 0): Concerning this last point I_3, the analytical analysis of stability is no more possible and it becomes then necessary to choose a parameter set.

Remark. Let's notice that the number of real positive fixed points depends on the choice of parameters.

5) Bifurcation analysis: Since I_2, J_2 and I_3 do not depend on parameters $(\beta_4, r_3, \alpha_1, \gamma_1)$, bifurcations can occur in these models (for both states of Colorado and Washington). Following the works of May [23], we propose to choose the parameter β_1, *i.e.*, the growth rate of the population of the prey (N) or the rate of moving in and out of the Nonuser category, as bifurcation parameter. Let's notice that this choice is based on the assumption according to which the population of Nonusers increases in proportion of the demography. Then, for the same reasons as previously, an analytical analysis would be difficult even impossible. So, in the next section we will set all the parameters except β_1 and then we will use a bifurcation diagram to determine the values of the bifurcation parameters.

6) Existence of bounded solutions: Following the works of Freedman [8] and while posing $N = x_1$, $E = x_2$, $R = x_3$ and $A = x_4$, the system (1) can be rewritten as follows:

$$\begin{aligned} \frac{dx_1}{dt} &= x_1 g(x_1) - (r_1 x_2 + \alpha_1 x_4 + \alpha_2 x_3) p_1(x_1), \\ \frac{dx_2}{dt} &= x_2 (-\beta_2 + r_1 p_1(x_1)) - (r_2 x_3 + \gamma_1 x_4) p_2(x_2), \\ \frac{dx_3}{dt} &= x_3 (-\beta_3 + r_2 p_2(x_2) + \alpha_2 p_1(x_1)) - r_3 x_4 p_3(x_3), \\ \frac{dx_4}{dt} &= x_4 [-\beta_4 + r_3 p_3(x_3) + \alpha_1 p_1(x_1) + \gamma_1 p_2(x_2)], \end{aligned} \tag{6}$$

where $g(x_1) = \beta_1(1 - x_1)$, $p_i(x_i) = \dfrac{x_i}{h + x_i}$ with $i = 1, 2, 3$.

Moreover, analysis of experimental data available on the prevalence of marijuana in the population of 21+ in the states of Colorado and Washington [15] has shown that the influence of A on N and E as well as that of R on N are in fact very weak. So, we will consider that $\alpha_1 \ll 1$, $\alpha_2 \ll 1$ and $\gamma_1 \ll 1$. Thus, under these assumptions, model (6) reads:

$$\begin{aligned} \frac{dx_1}{dt} &= x_1 g(x_1) - r_1 x_2 p_1(x_1), \\ \frac{dx_2}{dt} &= x_2 [-\beta_2 + r_1 p_1(x_1)] - r_2 x_3 p_2(x_2), \\ \frac{dx_3}{dt} &= x_3 [-\beta_3 + r_2 p_2(x_2)] - r_3 x_4 p_3(x_3), \\ \frac{dx_4}{dt} &= x_4 [-\beta_4 + r_3 p_3(x_3)]. \end{aligned} \tag{7}$$

Then, according to Theorem 2.1 stated by Freedman [8, p. 72], "all solutions initiating in the nonnegative cone are bounded and eventually enter a certain attracting set described below."

III. Applications of Predator-Prey NERA Model

In order to perform numerical experiments for forecasting the marijuana consumption in the states of Colorado and Washington beyond the year of the I – 502 implementation, we used data from Hanley [15]. The data collected in the latter report is the result of an analysis of population-level data in order to monitor four key indicators of marijuana use, namely, current marijuana use, lifetime marijuana use, marijuana abuse or dependency and age of initiation prior to implementation of I – 502. In the case of our NERA model, the first three categories correspond to the Recreational, Experimental and Addict category respectively [6]. The report highlights the importance of examining trends in that manner will allow them to monitor whether the implementation of I – 502 appears to affect these key indicators of marijuana use over time. Our numerical experiments aim to forecast the marijuana consumption in the states of Colorado and Washington beyond the year of the I – 502 implementation using data from Hanley [15]. The NERA predator-prey model is calibrated by estimating the parameters r_1, r_2, r_3, α_1, α_2,

α_3, β_1, β_2, β_3 and β_4. Parameter h has been arbitrarily chosen equal to $1/2$ as usually done in theoretical ecology [27]. To this aim, we used these data and a genetic algorithm as explained in Dauhoo *et al.* [6]. The fitness function used has been chosen to minimize the error that our model generates. MATLAB Optimtool is used and the fitness function is inserted in the genetic algorithm. We hence obtain the set of values in Tab. 2 and Tab. 3 corresponding to the NERA predator-prey model for Colorado and Washington. Obviously, since the parameters sets of both NERA predator-prey models have been calibrated starting from the data from Hanley [15], numerical integration of Eqs. (1) with parameters sets from Tab. II are in perfect agreement with the observed data. Thus, it confirms that the NERA predator-prey models can be a precious tool in forecasting of illicit drug consumption in the population of the state considered.

TABLE II: Parameter values for the consumption of Marijuana in Colorado and Washington

Parameter	Colorado	Washington
r_1	0.44	0.38
r_2	0.193	0.142
r_3	0.029	0.034
α_1	0.103	0.099
α_2	0.043	0.112
γ_1	0.031	0.032
β_1	0.042	0.015
β_2	0.016	0.03
β_3	0.052	0.066
β_4	0.047	0.039

A. *NERA model for Colorado*

Still using experimental data [15], we set the parameters from Tab. II where β_1 is the bifurcation parameter. By varying continuously β_1 from 0.02 to 0.8, we determine the values for which bifurcations occur by plotting the bifurcation diagram presented in Fig. 1.

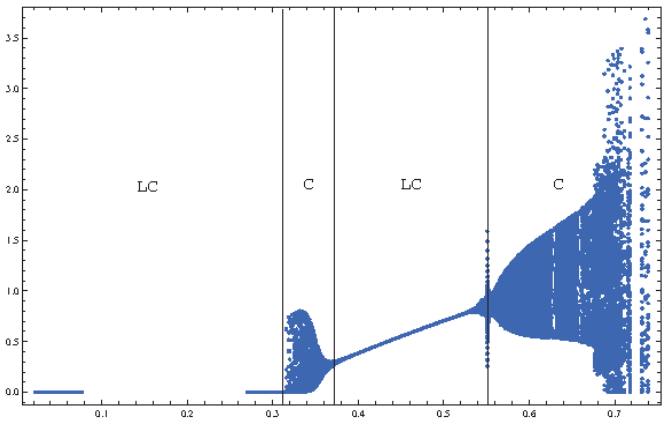


Fig. 1: Bifurcation diagram: u_{max} as function of β_1.

We observe in this diagram (Fig. 1), that for the value of $\beta_1 \in [0.02, 0.31]$ the solution of the NERA model (1) for Colorado is a *limit cycle* (LC) as exemplified in Fig. 2. Then, still increasing parameter β_1 up to the first bifurcation value $\beta_1^{b_1} = 0.3175$, the solution becomes chaotic (C) as highlighted in Fig. 3. Such chaotic feature persists up to the second bifurcation value $\beta_1^{b_2} = 0.37$ starting from which the solution becomes again a *limit cycle* (LC). Then, starting from the third bifurcation value $\beta_1^{b_3} = 0.55$ a chaotic attractor (C) similar to the previous one appears again.

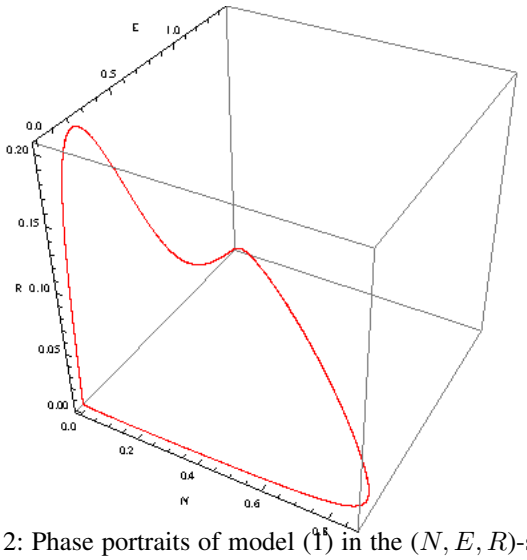


Fig. 2: Phase portraits of model (1) in the (N, E, R)-space for $\beta_1 = 0.3$.

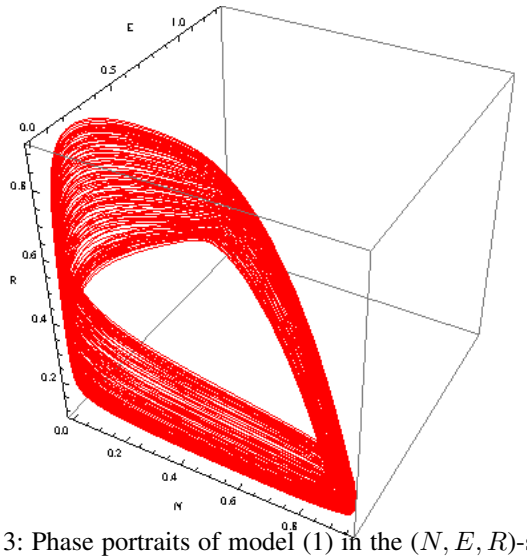


Fig. 3: Phase portraits of model (1) in the (N, E, R)-space for $\beta_1 = 0.35$.

The algorithm developed by Marco Sandri [26] for Mathematica® has been used to perform the numerical calculation of the Lyapunov characteristics exponents (LCE) of the NERA predator-prey model (1) for Colorado with the parameters from Tab. 2 and with $\beta_1 \in [0.02, 0.8]$. As an example for $\beta_1 =$

$0.3, 0.35, 0.4$ and 0.6, Sandri's algorithm has provided respectively the following LCEs $(0, -0.0022, -0.013, -0.029)$, $(0, 0, -0.0010, -0.065)$, $(0, -0.0011, -0.0011, -0.066)$ and $(0, 0, -0.0017, -0.085)$. Then, according to the works of Klein and Baier [19], a classification of (autonomous) continuous-time attractors of dynamical system (1) on the basis of their Lyapunov spectrum is presented in Tab. III. LCEs values have been also computed with the Lyapunov Exponents Toolbox (LET) developed by Pr. Steve Siu for MatLab® and involving the two algorithms proposed by Wolf *et al.* [35] and Eckmann and Ruelle [7] (see https://fr.mathworks.com/matlabcentral/fileexchange/233-let). Results obtained by both algorithms are consistent.

TABLE III: LCEs of NERA model (1) for Colorado for various values of β_1.

β_1	LCE spectrum	Dynamics of the attractor
$0.02 < \beta_1 < 0.31$	$(0, -, -, -)$	Periodic Motion (Limit Cycle)
$0.31 < \beta_1 < 0.37$	$(0, 0, -, -)$	3-Torus (Quasi-Periodic Motion)
$0.37 < \beta_1 < 0.55$	$(0, -, -, -)$	Periodic Motion (Limit Cycle)
$0.55 < \beta_1 < 0.8$	$(0, 0, -, -)$	3-Torus (Quasi-Periodic Motion)

B. NERA model for Washington

By varying continuously β_1 from 0.02 to 0.36 (all other parameters are those given in Tab. II), we determine the values for which bifurcations occur by plotting the bifurcation diagram presented in Fig. 4.

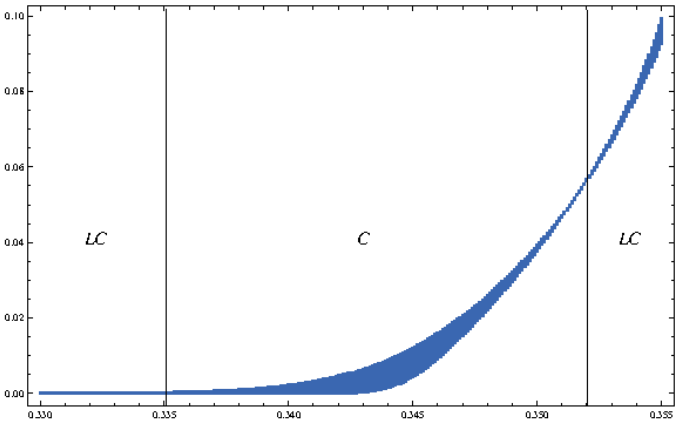


Fig. 4: Bifurcation diagram: u_{max} as function of β_1.

We observe from this diagram (Fig. 6) that for the value of $\beta_1 \in [0.02, 0.334]$ the solution of the NERA model (1) for Washington is a *limit cycle* (LC) in the (N, E)-plane. Then, still increasing parameter β_1 up to the first bifurcation value $\beta_1^{b_1} = 0.335$, the solution becomes chaotic (C). Such chaotic feature persists up to the second bifurcation value $\beta_1^{b_2} = 0.357$ starting from which the solution becomes again a *limit cycle*. Still using the algorithm developed by Marco Sandri [26] we numerically compute the Lyapunov characteristics exponents (LCE) of the NERA model (1) for Washington with the parameters from Tab. II and with $\beta_1 \in [0.02, 0.36]$. In this case, for $\beta_1 = 0.27, 0.34, 0.345$ and 0.3485, Sandri's algorithm provides respectively the following LCEs $(0, -0.0078, -0.0089, -0.019)$, $(0, 0, -0.010, -0.017)$, $(0, 0, -0.0092, -0.016)$ and $(0, 0, -0.011, -0.016)$. Then, as previously, a classification of attractors of dynamical system (1) on the basis of its Lyapunov spectrum is presented in Tab. 5. Here again, LCEs values have been also computed with the Lyapunov Exponents Toolbox (LET) and results obtained by both algorithms are consistent.

TABLE IV: LCEs of NERA model (1) for Washington for various values of β_1.

β_1	LCE spectrum	Dynamics of the attractor
$0.2 < \beta_1 < 0.33$	$(0, -, -, -)$	Periodic Motion (Limit Cycle)
$0.3375 < \beta_1 < 0.44$	$(0, 0, -, -)$	3-Torus
$0.44 < \beta_1 < 0.45$	$(0, -, -, -)$	Periodic Motion (Limit Cycle)

IV. Discussion

By considering that drug users are classified into four main categories: non (N), experimental (E), recreational (R) and addicts (A) users, Dauhoo *et al.* [6] have proposed the NERA model. Nevertheless, although this four-dimensional dynamical system took into account the mutual influence that drug users (E, R and A) can have on non-users (N) and on each other, it did not contain any limitation in the growth and decay of each variable. Thus, no oscillatory or chaotic regime could be observed. So, the aim of this work was to propose a modified version of this NERA model by analogy with the classical predator-prey models and while considering non-users (N) as prey and users (E, R and A) as predator. Thus, this new model included a "stabilizing effect" of the growth rate of the preys (N) and a "destabilizing effect" of the predators (E, R and A) saturation. Functional responses of Verhulst and Holling type II have been used for modeling these effects. Then, in order to perform numerical experiments for forecasting the marijuana consumption in the states of Colorado and Washington beyond the year of the I – 502 implementation, we used data from Hanley [15] and a genetic algorithm as explained in Dauhoo *et al.* [6]. Thus, the NERA predator-prey model has been calibrated by estimating all the parameters except h which has been arbitrarily chosen equal to $1/2$ as usually done in theoretical ecology [27]. We hence obtained two parameters sets corresponding to the NERA predator-prey model for the state of Colorado and Washington. Following the works of May [23], we chose the parameter β_1, *i.e.*, the growth rate of the population of the prey (N) or the rate of moving in and out of the Nonuser category, as bifurcation parameter. This choice was based on the assumption that the population of Nonusers increases in proportion of the demography. A stability and bifurcation analysis of the NERA model for Colorado and for Washington has therefore been performed.

Concerning the NERA model for Colorado, the bifurcation diagram has shown that when the value of parameter $\beta_1 \in [0.02, 0.31]$, the solution is a *limit cycle* confirming thus the behavior of the observed data from Hanley [15]. So, the number of individuals of each group N, E, R and A oscillates in a deterministic way with a period and amplitude that can be numerically computed. Then, still increasing parameter β_1 up to the first bifurcation value $\beta_1^{b_1} = 0.3175$, we have shown that the solution becomes quasi periodic (3-Torus). Such chaotic attractor persists up to the second bifurcation value $\beta_1^{b_2} = 0.37$ starting from which the solution becomes again a *limit cycle*. When parameter $\beta_1^{b_3} = 0.55$ reaches the third bifurcation value a chaotic attractor appears again. These results have been confirmed by the computation of the Lyapunov characteristics exponents.

Concerning the NERA model for Washington, the bifurcation diagram has shown that when the value of parameter $\beta_1 \in [0.02, 0.26]$ the solution is a *limit cycle* confirming thus the behavior of the observed data from Hanley [15]. Then, still increasing parameter β_1 up to the first bifurcation value $\beta_1^{b_1} = 0.34$, the solution becomes quasi periodic (3-Torus). Such chaotic attractor persists up to the second bifurcation value $\beta_1^{b_2} = 0.357$ starting from which the solution becomes again a *limit cycle*. These results have been also confirmed by the computation of the Lyapunov characteristics exponents.

V. Conclusion

Thus, we have shown that an increase of the population of non-users (N) leads first to a periodic evolution of drug-users (E, R, A). But when this increase reaches a certain threshold for both states of Colorado and Washington, the evolution of all variables becomes unpredictable. Such result can be also interpreted by analogy with the so-called "paradox of enrichment" which states that increasing the food available to the prey caused the predator's population to destabilize [25]. So, we have confirmed on one hand that the NERA predator-prey models can be a precious tool in forecasting of illicit drug consumption in the population of the state considered and, on the other hand that they can be of substantial interest to policy-makers in the problematic of illicit drug consumption.

References

[1] A. Baveja, R. Batta, J. P. Caulkins & M. H. Karwan, *Modeling the response of illicit drug markets to local enforcement*, Socio-Economic Planning Sciences, 27(2), (1993) p. 73-89.

[2] A. Baveja, R. Batta, J. P. Caulkins & M. H. Karwan, *Collapsing street markets for illicit drugs: the benefits of being decisive*. Working Paper 93-39, Heinz School of Public Policy and Management, Carnegie Mellon University, Pittsburgh, PA (1993).

[3] J. P. Caulkins (Ed.), *Mathematical Models of Drug Markets and Drug Policy*, Special issue of Mathematical and Computer Modelling 17, (1993) p. 1-115.

[4] J. P. Caulkins, *Local drug markets' response to focused police enforcement*, Operations Res. 41(5), (1993) p. 848-863.

[5] J. P. Caulkins, R. C. Larson & T. F. Rich, *Geography's impact on the success of focused local drug enforcement operations*, Socio-Economic Planning Sci. 27(1), (1993) p. 119-130.

[6] M. Dauhoo, B. Korimboccus & S. Issack, *On the dynamics of illicit drug consumption in a given population*, IMA Journal of Applied Mathematics, 78(3), (2013) p. 432-448.

[7] J.P. Eckmann & D. Ruelle, *Ergodic Theory of Chaos and Strange Attractors*, Rev. Mod. Phys., 57, (1985) p. 617-656.

[8] H.I. Freedman & J.W.H.So, *Global Stability and Persistence of Simple Food Chains*, Mathematical Biosciences, **76**(1) (1985) 69–86.

[9] G.F. Gause, *The struggle for existence*, Williams and Wilkins, Baltimore, (1935).

[10] G.F. Gause, *Vérifications expérimentales de la théorie mathématique de la lutte pour la vie*, Hermann, Paris, (1935).

[11] J.M. Ginoux, *The paradox of Vito Volterra's predator-prey model*, Lettera Matematica, Springer-Verlag, Milan (2017) p. 1-7.

[12] A. Gragnani, G. Feichtinger & S. Rinaldi, *Dynamics of drug consumption: a theoretical model*, International Insitute for Applied Systems Analysis, Laxenburg, Austria, Working Paper 94-77 (1994).

[13] A. Gragnani, S. Rinaldi & G. Feichtinger, *Slow-fast limit cycles in controlled drug markets*, Proceedings of 3rd European Control Conference, Rome, Italy, Vol 4 (1995) p. 3031-3034.

[14] A. Gragnani, S. Rinaldi & G. Feichtinger, *Dynamics of drug consumption: a theoretical model*, Socio-Economic Planning Sciences, 31(2), (1997) p. 127-137.

[15] S. Hanley, *Legalization of Recreational Marijuana in Washington: Monitoring Trends in Use Prior to the Implementation of I-502*, http://www.wsipp.wa.gov/ReportFile/1540/Wsipp Legalization-of-Recreational-Marijuana- in-Washington-Monitoring-Trends-in-Use-Prior-to-the-Implementation-of-I-502 Full-Report.pdf.

[16] C.S. Holling, *The components of predation as revealed by a study of small-mammal predation of the European pine sawfly*, Can. Entomol. 91, (1959) p. 293-320

[17] C.S. Holling, *Some characteristics of simple types of predation and parasitism*, Can. Entomol. 91, (1959) p. 385-398.

[18] E. Hopf, *Abzweigung einer periodischen Lösung von einer stationären Lösung eines Differentialsystems*, Berichten der Mathematisch-Physischen Klasse der Sächsischen Akademie der Wissenschaften zu Leipzig XCIV, (1942) p. 1-22 (1942). An English translation, with comments, is included as Section 5 in Marsden *et al.* [1976].

[19] M. Klein and G. Baier *Hierachies of dynamical systems*, in A Chaotic Hierarchy, Singapore: World Scientific (1991).

[20] Y.A. Kuznetsov, *Elements of Applied Bifurcation Theory*, Springer-Verlag, Berlin, 1998.

[21] J.E. Marsden & M. McCracken, *The Hopf Bifurcation and Its Applications*, Springer-Verlag, New York, (1976).

[22] R.M. May, *Theoretical Ecology: Principles and Applications*, Saunders, New York, (1976).

[23] R.M. May, *Simple mathematical models with very complicated dynamics*, Nature 261, (1976) p. 459-467.

[24] H. Poincaré, H. *Méthodes Nouvelles de la Mécanique Céleste*, Paris : Gauthier-Villars, volume I, II & III (1892-93-99).

[25] M. Rosenzweig, *The Paradox of Enrichment*, Science 171, (1971), p. 385-387.

[26] M. Sandri *Numerical Calculation of Lyapunov Exponents*, The Mathematica Journal, 6(3), (1996) p. 78-84.

[27] F.M. Scudo & J.R. Ziegler, *The Golden Age of Theoretical Ecology 1923-1940*, Lecture Notes in Biomathematic, vol. 22, Springer-Verlag, Berlin-Heidelberg-New York 1978.

[28] L.P. Shilnikov L.P. *A case of the existence of a denumerable set of periodic motions*,Sov. Math. Dokl., 6 (1965), p. 163-166.

[29] L.P. Shilnikov *The existence of a denumerable set of periodic motions in four-dimensional space in an extended neighborhood of a saddle-focus*, Sov. Math. Dokl., 8(1) (1967), p.54-58.

[30] P.F. Verhulst, *Notice sur la loi que suit la population dans son accroissement*, Corresp. Math. Phys., **X** (1838) 113–121.

[31] V. Volterra, "Variazioni e fluttuazioni del numero d'individui in specie animali conviventi," Mem. Acad. Lincei III, 6 , 31–113 (1926).

[32] V. Volterra, *Fluctuations in the abundance of a species considered mathematically* ("Nature", vol. CXVIII, 1926_2, pp. 558—560). Sotto lo stesso titolo furono poi pubblicate due lettere, una del LOTKA e una del VOLTERRA (Ibidem, vol. CXIX, 1927, pp. 12–13).

[33] V. Volterra, *Variation and fluctuations of the number of individuals in animal species living together*. Translated by Miss MARY EVELYN WELLS ("Journal du Conseil international l'exploration de la mer", Copenhague, vol. III, n. I, 1928, pp. 3–51).

[34] V. Volterra, *Leçons sur la Théorie Mathématique de la Lutte pour la Vie*, Gauthier-Villars, Paris, (1931).

[35] A. Wolf, J.B. Swift, H.L. Swinney & J.A. Vastano, *Determining Lyapunov Exponents from a Time Series*, Physica D, 16, (1985) p. 285-317.

Proficient Algorithm for Features Mining in Fundus Images through Content Based Image Retrieval

C. Gururaj
(IEEE Member)
Dept. of Telecommunication Engg.
BMS College of Engineering
Bengaluru, Karnataka, India
gururaj.tce@bmsce.ac.in

***Abstract*—Image retrieval is gaining prominence in the area of medical image processing especially in the domain of fundus images. This work aims to propose a proficient algorithm for features mining in Fundus images and thereby extract the information through Content Based Image Retrieval process. The automated extraction of important features such as exudates aids medical practitioners in effectively overcoming various diseases pertaining to the patient. Although multiple methods of extracting these features are available, they lack in retrieval aspect of the information or the accuracy of the feature extraction.**

Keywords—CBIR, feature extraction, fundus images, diabetic retinopathy, histogram

I. INTRODUCTION

In today's world, the searchable picture information internally exhibits the various visual and the semantic substance; also fluctuating geographic areas are quickly developing in size and limit [1]. These variables have made endless potential outcomes and thus contemplations for appearance investigation framework fashioners. However, genuine utilization of the innovation is as of now constrained. This area is dedicated for seeing twofold recuperation that clarifies client's desires, additionally the association choking influences likewise framework necessities, and the exertion taken for the examination to make duplicate recuperation a sensible and material one [2].

Images related to architecture, medicine and engineering are increasing fourfold each day and making use of the information embedded in each image is a great challenge [4]. The CBIR technique can be based on basic image features including but not limited to colour, shape and texture [5]. This research deals with discussion of popular methods of performing CBIR. The role of CBIR in the domain of medicine is also discussed in this work. The advantage of this work is that it aids the medical practitioners in effective identification of syndrome in the quickest possible duration [3]. The details of the afflictions affecting the eye are discussed in detail in subsequent sections.

All the important regions of the internal eye constitute the 'fundus' region of the eye and it is found exactly opposite to the lens of the eye. The fundus region includes the fovea, retina, macula, optic disc, posterior pole [6] and also the blood vessels present in the eye. The ophthalmoscope can be used to inspect the fundus region of the eye [7]. This is done by placing the ophthalmoscope near the eye supported with bright illumination pointed towards the eye. The eye is initially dilated before it undergoes this process. The diagnosis of the fundus region can also be done through a 'Fundus camera'. The fundus camera works on the same principle as that of the ophthalmoscope and consists of a camera of high resolution as well. Several medical conditions of the eye such as exudates, abnormalities and constrictions in the blood vessels, haemorrhages, cotton wool spots etc. can be observed with the help of these images procured from the fundus camera.

According to "World Health Organization" (WHO) the survey made on screening for diseases specifies that in the year 2010 globally around millions of people visually had become blind. Based on these estimation 80% cases for visual blindness are preventable or treatable. Diabetic Retinopathy (DR) and Age related Macular Degeneration (AMD) is two most frequently observed diseases which lead to loss of a visualization as shown in Fig. (1) [8]. Thus immediate and faster diagnosis will help for a better treatment, which reduces the cost when compared in advanced phases which may later also become severe.

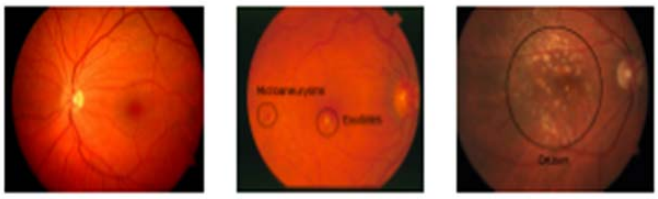


Fig. 1. Fundus Images (a)Healthy (b) DR (c) AMD

The Content Based Image Retrieval (CBIR) helps in ordering comparably sparing images of the source visual substance. CBIR fathoms a few costs that are connected through old methodologies of recovering images by its catchphrases and comments. The introduction of the CBIR arrangement relies upon specific image portrayal and similitude coordinating strategies, which are discussed in the subsequent sections.

II. CBIR APPROACHES AND PERFORMANCE METRICS

A. CBIR Approaches

Distinctive procedures are accessible for the strategy of extraction and access of image annotations delivered on the histograms like nearby or worldwide, shading formats, edges, slopes, shapes, surfaces and shapes, limits and districts that have been investigated in the above writing [9], [10]. The current improvements that are made in the field of CBIR comprise of Relevance Feedback, Radial base occupation, fluffy procedures and so forth. These strategies include arbitrary routine of human cooperation in a material

way, in this way it essentially diminishes yield recovery time.

Relevance Feedback (RF) [11] is a strategy used to tackle the issue called semantic crevice, in which the clients give input on the importance of the recovered pictures by the framework, and this data is offered back to the framework for taking in the client data needs.

Radial basis function (RBF) systems are generally utilized for example arrangement. RBF organizes commonly have two particular layers. The lowest level contains a utilization of essential capacities each of which is a motivation behind the separation between an info design and a model example.

By and large, Clustering calculation [12] are the real wellspring of prevalent association found in utilizing a multidimensional information as it gives a compelling recovery grouping strategy and thoughts that permits the element seek procedure of KNN. Utilizing fluffy calculation, the term set up can be produced and the purpose of Fuzzy rationale contracts utilizing the thinking and breaking down process, which is a rough procedure related towards the settled and correct process.

Neural network is set of associated neurons. This is utilized for whole guess. ANN [13] are typical to contain many interconnecting simulated neurons and systems, which helps in giving the information required to the comprehension of natural neural systems or essentially to solve the Artificial Intelligence (AI) issues.

The Support Vector Machine (SVM) is the best example in class grouping technique. The classifier of SVM is prevalently utilized as a part of the application and investigation of bioinformatics and numerous different teaches because of its superior in exactness, capacity of preparing the high dimensional information issues.

The fundamental point of K-mean bunching technique [14] is to arrange the information made on its bit detail strategies. It makes a particular level of group's diverse ordered clustering strategy's tree structure. The measure of separation decides the examination between the components found in the framework.

B. Algorithm Performance Metrics

Different performance metrics are used to authenticate the effectiveness of the research work. Specifically the performance metrics considered in this work are MSE and PSNR. These metrics ensure the efficacy of the algorithm.

The deviation between the obtained result and the ideal estimated output is given through the MSE. The quality of the output image files is measured through PSNR.

III. FEATURE EXTRACTION USING HISTOGRAM ANALYSIS

Conventional Colour Histogram (CCH) [CG] strategy is all the more constantly grasped for CBIR frameworks CCH. Here, the pixel of the separate pictures is identified with an unequivocal histogram container on the base of shading, similitude in the canister. While executing with the numerous receptacles, the computational many-sided quality is expanded with effective building list. Apparently, quantization diminishes the data in light of the substance of pictures however the appealing part of CCH makes the operation straightforward and simplicity of calculation.

Invariant Colour Histogram is widely used for protest acknowledgment. However, these histograms are varying steadily under the varieties in perspective, it clears up that the shading histogram created from a picture surface is firmly identified with the geometry surface, and the review position. A plan is built up to create shading histogram, which relies on upon the shading slopes and it is invariant under mapping surface that is locally relative.

For the most part, two steady focuses in two pictures, one can change sort out frameworks with the end goal that both focuses lie at the birthplace in their particular frameworks through Local relative approximations. Expecting that the modifications among the two pictures are nonstop, it is locally straight. This is determined as a feeble supposition, fulfilled by changes, for example, homographs, alterations in view-point or smooth misshapenness. Consequently, for microscopic x the nearby relative change as shown in (1)

$$H = HX' \tag{1}$$

where $H = \begin{bmatrix} AB \\ CD \end{bmatrix}$

Here, x' and x are the focuses in the first and second arrange frameworks, separately.

Differential range from angles by definition is given in (2)

$$da' = dx'dy' \tag{2}$$

At that point locate the comparing range in the second facilitate framework using (3)

$$da = \left|\frac{\partial X}{\partial x'} \times \frac{\partial X}{\partial y'}\right| = |AD - BC|dx'dy' \tag{3}$$

Outline the Image Intensity Function (IIF) in one shading channel by using (4)

$$f(x, y) = f'(x', y') \tag{4}$$

Utilizing standard mathematics, we can relate the picture subsidiaries as shown from (5) through (8)

$$\frac{\partial f'}{\partial x'} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial x'} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial x'} \tag{5}$$

$$\frac{\partial f'}{\partial y'} = \frac{\partial f}{\partial x}\frac{\partial x}{\partial y'} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial y'} \tag{6}$$

$$f_x' = f_x A + f_y C \tag{7}$$

$$f_y' = f_x B + f_y D \tag{8}$$

Exhibiting the IIF in moment shading by a condition in (9), and (10). The subordinates will comply with indistinguishable imperatives from above conditions (7) and (8).

$$g_x' = g_x A + g_y C \quad (9)$$

$$g_y' = g_x B + g_y D \quad (10)$$

Presently, given all subsidiaries, it is conceivable to fathom for each of A, B, C, and D as specified in condition (11) through (14).

$$A = \frac{f_y g_x' - g_y f_x'}{f_x g_y - g_x f_y} \quad (11)$$

$$B = \frac{f_y g_y' - g_y f_y'}{f_x g_y - g_x f_y} \quad (12)$$

$$C = \frac{f_c g_x' - g_c f_x'}{f_x g_y - g_x f_y} \quad (13)$$

$$D = \frac{f_c g_y' - g_c f_y'}{f_x g_y - g_x f_y} \quad (14)$$

Hence,

$$da = |AD - CB| = \frac{f_x g_y' - f_y' g_x}{f_x g_y - f_y g_x} \quad (15)$$

It assesses the immediate association among the shading subsidiaries, and the differential zones.

$$da\left|f_x g_y - f_y g_x\right| = da'\left|f_x g_y' - f_y' g_x\right| \quad (16)$$

The accompanying condition (16) is the key relationship that is used to develop invariant shading histograms. Utilizing the above conditions, a straightforward calculation to make twisting invariant shading histograms is accomplished.

The above mathematical modelling forms the core of the algorithm and based on the same, the proposed block diagram is shown in Fig. (2). Image enhancement is an integral and important step that is considered to obtain additional information of the image and also to preserve the basic features of the image [15]. Also the enhancement methodology is based on improving two important parameters i.e. illumination and contrast variation.

When changes are applied on channels other than the luminous channel, the basic challenge of colour distortion is to be considered. Hence the luminosity channel is considered for the process. The colour space considered is that of HSV which in turn is used to amplify the RGB channels respectively. Adaptive histogram equalization [1] is used to augment the contrast of the luminosity channel in the L*a*b colour space. This step aids in overcoming the heterogeneous regions in the fundus image.

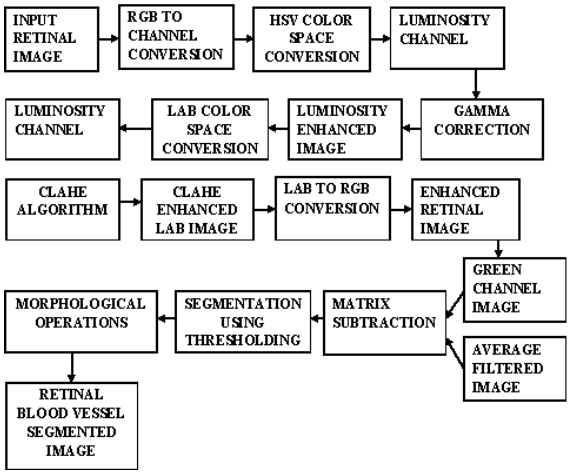


Fig. 2. Proposed block diagram

IV. RESULTS

The retinal images in the DRIVE dataset [8] were considered for the proposed algorithm. It may be noticed from the results in Table I through III that algorithm performs well when enhanced images are considered. Some of the other parameters considered for analysis of the algorithm include the MSE and PSNR.

One of the major challenges during the implementation of the algorithm was the detection of the optic disk along with other afflictions due to inherent high luminosity value. This generally gives rise to false positive detection of lesions. Hence separation of the optic disk from actual afflictions is also taken care in this work. The present work extracts the blood vessels from a retina image using the adaptive histogram equalization algorithm. The smoothened image is subtracted with the enhanced image. From that, the blood vessel region will be highlighted. Then we apply thresholding algorithm and then segment the blood vessel region. To remove the unwanted objects, we are further using morphological operations.

The results at various stages of processing the image is captured as follows:

Fig. 3. Input Image

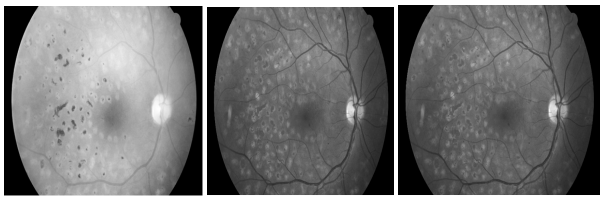


Fig. 4. R, G and B components of Original Image

Fig. 5. Background removal for R G B

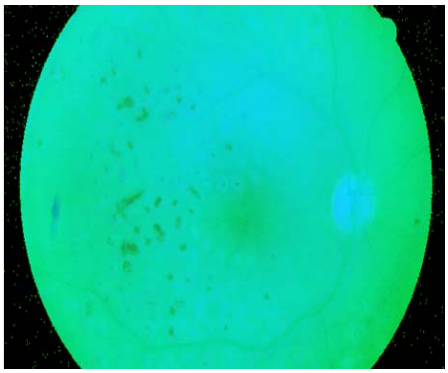


Fig. 6. H S V Image

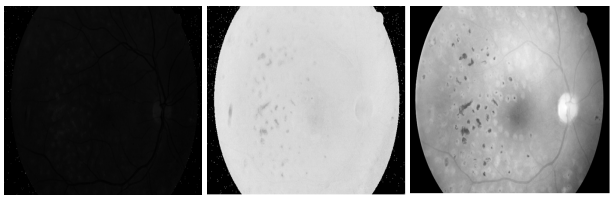


Fig. 7. Hue, Saturation and Value channels of the Image

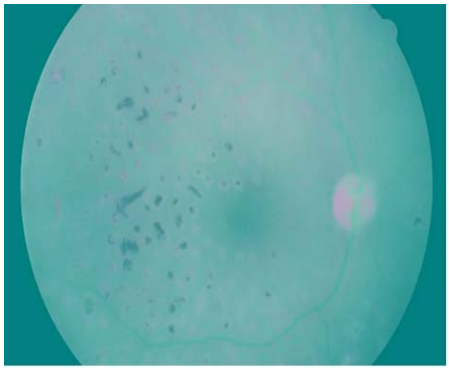


Fig. 8. Lab Colorspace Image

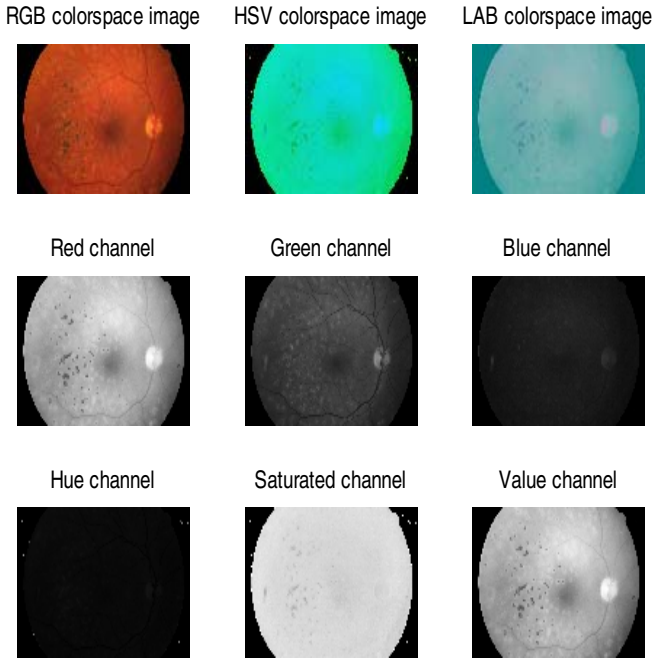


Fig. 9. R G B, H S V and Lab Colorspace Images

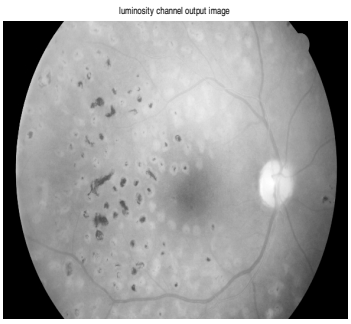


Fig. 10. Luminosity Channel output Image

Fig. 11. Gamma Correction applied R G B Image

Fig. 12. Contrast Enhanced R G B Image

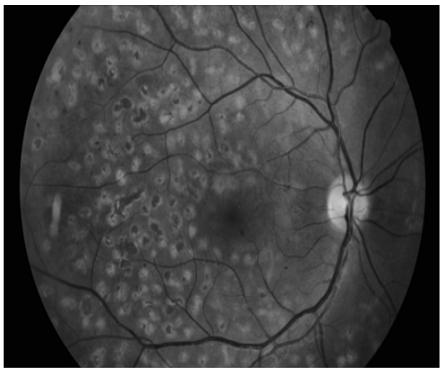


Fig. 13. Average Filter Image

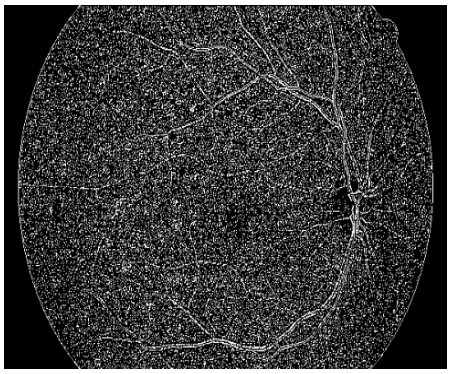


Fig. 14. Binarized Image

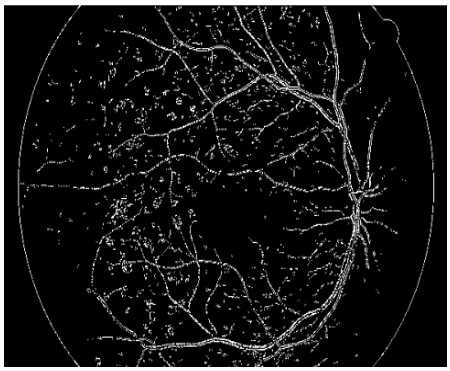


Fig. 15. Vessel Segmented Image

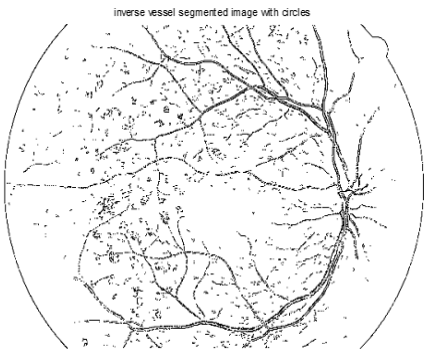


Fig. 16. Inverse Vessel Segmented Image

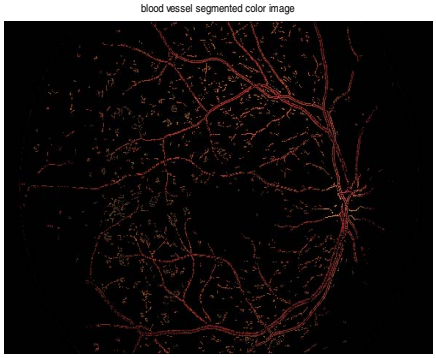


Fig. 17. Blood Vessel Segmented Image

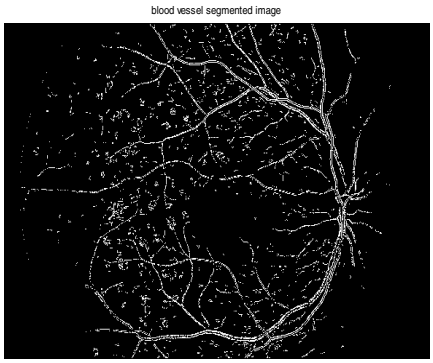


Fig. 18. Binary Blood Vessel Segmented Image

TABLE I. INPUT IMAGE ANALYSIS

Sl.no	*Mean of input*	*SD*
1	0.3370	0.3434
2	0.3082	0.3019
3	0.3804	0.3177
4	0.2839	0.2632
5	0.3220	0.3152
6	0.3168	0.3112
7	0.2872	0.2581
8	0.4015	0.3144
9	0.3416	0.2686
10	0.2799	0.2892

TABLE II. L*A*B ,H S I AND H S V IMAGE ANALYSIS

Sl.no	*MSE*	*PSNR*	*Mean of input*	*SD*
1	0.0027	75.1351	0.4611	0.2672
2	0.0042	72.9816	0.4449	0.2621
3	0.0016	74.5938	0.5073	0.2854
4	0.0096	69.4103	0.4368	0.2628
5	0.0041	73.0627	0.4621	0.2824
6	0.0036	71.4846	0.4697	0.2846
7	0.0126	68.1925	0.4421	0.2586
8	0.0016	75.1051	0.5347	0.2933

Sl.no	*MSE*	*PSNR*	*Mean of input*	*SD*
9	0.0076	69.3102	0.4922	0.2589
10	0.0082	67.9403	0.4243	0.2716

TABLE III. ENHANCED IMAGE ANALYSIS

Sl.no	*MSE*	*PSNR*	*Mean of input*	*SD*
1	0.0027	77.1263	0.4611	0.1871
2	0.0042	73.9824	0.4449	0.1962
3	0.0016	75.5938	0.5073	0.2846
4	0.0096	69.9132	0.4368	0.2620
5	0.0041	73.1423	0.4621	0.2823
6	0.0036	72.3891	0.4697	0.2832
7	0.0126	69.2619	0.4421	0.2487
8	0.0016	75.8053	0.5347	0.2922
9	0.0076	69.9133	0.4922	0.2576
10	0.0082	68.2403	0.4243	0.2711

V. CONCLUSIONS

A simple and effective system for image enhancement is suggested that includes various enhancement techniques including enhancements of both luminosity and contrast. From the results it can be observed that the proposed algorithm performs in a superior manner compared to other approaches for the color retinal images. The proposed work would aid ophthalmologists towards an effective diagnosis of afflictions in the eye. The algorithm could also be used towards implementation of a standalone automated system that achieves the retrieval process.

ACKNOWLEDGMENT

The author acknowledge the support of the management of BMSCE for supporting the work reported in this paper.

REFERENCES

[1] Ten Brinke, Walter, David McG, and John Bigelow. "The Meaning of an Image in Content-Based Image Retrieval." PhiSE. 2006.

[2] Datta, Ritendra, "Image retrieval: Ideas, influences, and trends of the new age." ACM Computing Surveys (CSUR) 40.2 (2008): 5.

[3] Mei Zhou, Kai Jin, Shaoze Wang, Juan Ye, Dahong Qian. "Color Retinal Image Enhancement Based on Luminosity and Contrast Adjustment", IEEE Transactions on Biomedical Engineering, pp521-527. Vol. 65, No. 3, March 2018

[4] Weidong Cai et.al, "Content-based retrieval of dynamic PET functional images" IEEE Transactions on Information Technology in Biomedicine, 4(2): 152-158, June 2000.

[5] Kotoulas, L. and Andreadis I., "Colour histogram content-based image retrieval and hardware implementation", IEEE Proceedings on Circuits, Devices and Systems, 150(5): 387-393, October 2003.

[6] Asim, K.M.; Basit, A.; Jalil, A., "Detection and localization of fovea in human retinal fundus images," Emerging Technologies (ICET), 2012 International Conference on , vol., no., pp.1-5, 8-9 October 2012.

[7] Pereira T. .; Barbeiro, P.; Lemos, J.; Morgado, M.; Silva, E., "Digital image acquisition for ophthalmoscope," Bioengineering (ENBENG), 2012 IEEE 2nd Portuguese Meeting in , vol., no., pp.1-6, 23-25 Feb. 2012.

[8] J.J. Staal, M.D. Abramoff, M. Niemeijer, M.A. Viergever, B. van Ginneken, "Ridge based vessel segmentation in color images of the retina", IEEE Transactions on Medical Imaging, vol. 23, pp. 501-509, 2004.

[9] Singh, Amit, Parag Sohoni, and Manoj Kumar. "A Review of Different Content Based Image Retrieval Techniques."International Journal of Engineering Research and General Science Volume 2, Issue 5, August – September 2014.

[10] O K Girija, M.SudheepElayidom "Recent Trends in Image Retrieval Techniques for the Big Data Platform: A Survey", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, Special Issue 1, Feb. 2016.

[11] Sundaresan, S.M. and Srinivasagan, D. K, "Design of Image Retrieval Efficacy System Based on CBIR", Int. J. of Advanced Research in Computer Science and Software Engineering, 3(4), pp.48-53, 2013.

[12] H. P. Ng, S. H. Ong, K. W. .Foong, P. S. Goh, and W. L. Nowinski, "Medical Image Segmentation Using K-Means Clustering and Improved Watershed Algorithm," IEEE Southwest Symposium on Image Analysis and Interpretation, pp. 61-65, 2006.

[13] C.Gururaj, D Jayadevappa, Satish Tunga, "Content Based Image Retrieval System Implementation through Neural Network", IOSR Journal of VLSI and Signal Processing (IOSR-JVSP), Volume 6, Issue 3, Ver. 3, e-ISSN: 2319 – 4200, p-ISSN No. : 2319 – 4197, pp 42 – 47, May - June 2016.

[14] C. W. Chen, J. Luo, and K. J. Parker, "Image segmentation via adaptive K-mean clustering and knowledge-based morphological operations with biomedical applications," IEEE Transactions on Image Processing, vol. 7, no. 12, pp. 1673-1683, Dec. 1998.

[15] C Gururaj, D Jayadevappa, Satish Tunga, "Fundus Image Features Extraction for Exudate Mining in Coordination with Content Based Image Retrieval: A Study", Journal of The Institution of Engineers (India): Series B, Springer-Verlag, Published online: 5th Feb 2018, volume 99, Issue 3, pp 313-321, ISSN(print): 2250-2106, ISSN(Electronic): 2250-2114, DOI: 10.1007/s40031-018-0319-7

The Potential of Second Life as a Platform for Learning: Student's Perspective

Aphiwe Linganisa
Discipline of Information Systems & Technology
University of KwaZulu-Natal
Pietermaritzburg, South Africa
lingania80@gmail.com

Anyetei Ako-Nai
Discipline of Information Systems & Technology
University of KwaZulu-Natal
Pietermaritzburg, South Africa
akonaia@ukzn.ac.za

Nurudeen Ajayi
Discipline of Information Systems & Technology
University of KwaZulu-Natal
Pietermaritzburg, South Africa
ajayi@ukzn.ac.za

***Abstract*—Educational institutions of today are increasingly adopting various technologies, so as to enhance their teaching and learning activities. Second Life (SL) is one of such technologies being employed by educational institutions to create an environment that facilitates learning. SL is believed to facilitate engagement and conceptualization of knowledge. Its adoption in education and its enablement of learning has, however, been faced with challenges that require the understanding of its usefulness as a learning platform. This study, therefore, investigated the potential of SL as a learning platform, by investigating the perception of students on its usefulness as a learning platform. The study also investigated the challenges that come with the use of SL as a learning platform. The qualitative research method and the Technology Adoption Model (TAM) were adopted in this study. Using purposive sampling, the sample was drawn from amongst fourth-year information systems and technology (IS&T) students at the University of KwaZulu-Natal (UKZN). Findings of the study reveal that SL is a useful learning platform and students identified it as a beneficial platform that improves their ability to be creative in their learning. Students, however, also identified the challenges of using SL to include inadequate training and insufficient information technology (IT) infrastructure. To address these challenges, they suggested adequate training and the provision of infrastructure to support the adoption of SL as a learning platform.**

Keywords—Second Life, Virtual World, Technology, TAM, Teaching and Learning

I. INTRODUCTION

Technology has been changing rapidly over the years, and as a result, has presented institutions with a number of learning platforms that can be used to facilitate teaching and learning. Virtual Reality (VR) is one of the recently added technology believed to have the potential to transform educational processes [1]. VR technology and immersive environments have become more practical and useable, as seen in the case of modern video games and television shows [2]. They are also increasingly becoming vital tools in the facilitation of student-centered learning and in the promotion of flexible and distance learning. One of the prominent VR technologies that is being increasing used to facilitate learning is second life (SL). According to Saeed, Yang [3], SL extends the functionalities of other technologies by creating a dynamic virtual environment that provide students with real-time learning experience. The use of SL to facilitate learning is, however, faced with challenges such as understanding and benefiting from its usefulness as a learning platform. This study therefore tries to understand, from the perspective of students, the usefulness of SL as a learning platform. In the next section, the paper presents the literature review on Virtual World (VW) and Second Life, and their relevance in education. This is then followed by the theoretical framework used as the bases of the study, then by the methodology. The study findings and the conclusion on the study are the last sections of the paper.

II. LITERATURE REVIEW

A. Virtual World

VW is an environment that simulates real or fictional life [4], and that provides a phenomenal experience and a common space for users to engage in online activities [5]. In the study of Warburton [6], it was described as a computer-generated environment that give users a sense of being present in an environment (virtual), when in actual sense, they are not present (i.e. physically) in the environment. It was also described in the study of Dreher, Reiners [7] to be a web-based application that allow user(s) to be present and also be involved in an online game-like environment, but without game-like rules. In VWs, users interact with each other through their virtual representation known as "*avatar*" [8], which according to Loke [4], are the graphical representation of users within the VW.

According to Warburton [6], VWs have existed since the early 1980s. However, they have evolved from being mainly a gaming or social interaction platform to a platform where organizations have optimized opportunities by gaining from subscription fees, advertising opportunities or offering support services. They are also now increasingly being used for business development and for conducting meetings and conferences, where users can collaborate effectively despite being in different physical locations. Furthermore, in educational environments, it is becoming a potent tool for conducting research because it can accommodate the recruitment of a wide-range of different people from different backgrounds to take part in interviews, experiments, focus groups or surveys.

B. Second Life

SL is a content creation type of VW where users, represented as avatars, can par-take in various activities and challenges as part of their learning experience. It provides a platform for users to 'live a second life', though in a virtual environment [9]. In this life, residents are able to build houses, buy cars, develop businesses and relationships and even own the intellectual property (IP) right to their virtual properties. They are also able to develop environments (virtual) in which they can make gestures such as walking, dancing and teleporting. According to Palomäki [10], SL provides a culturally diverse experience and a livelier communication because it enable users to customize their

avatar to suit their culture, race or religion while meeting and collaborating with people that are from different background.

The main characteristics of SL is its game-oriented interfaces and its user-friendly scripting language [8]. Residents in SL, through their avatars, can interact with the game-oriented interface of SL, and can communicate with each other through voice and chat messages. The voice and chat functionalities provided in the SL environment makes it easy for shy users to socialize, such that they do not have to feel as intimidated as they would if they were socializing in the real world [11]. According to Falloon [12], SL provides a virtual space where teachers and students can actively immerse themselves in the learning process and in turn, develop effective teaching and learning skills. In the virtual space, students actively engage in various learning activities with clear learning objectives, as set out by the teacher.

C. *Second Life in Education*

Modern teaching and learning should focus not only on the environment in which cognitive learning takes place, but also, on the role of students in being creative and in constructing their own knowledge. According to Dreher, Reiners [13], the constructive and creative way of learning is facilitated and driven by SL. Thus, making SL a platform suitable for adoption in education. SL is increasingly being used for educational purposes because of the features it offers, which has not only brought enhancement in the learning experience of students, but has also provided a positive shift away from the traditional way of learning. In the study of Warburton [6], SL was identified as one of the most popular VW platform being used in education. In the study it was shown that the adoption of SL in education makes it easy for students to visualize and engage with the concepts they are being taught. In the study of Ali, Ahmed [14], it was also shown that the adoption of SL in education promotes experiential learning, in which students are able to interact in a realistic environment, that supports the building of new knowledge through their own experiences.

Furthermore, the study of Esteves, Fonseca [15], shows that using SL platform as an educational tool provides an immersive learning environment where simulations and experiments can be done in a safe environment. According to Olasoji and Henderson-Begg [1], the SL environment is considered safe because it eliminates the risks involved in the real world, for example, in the simulation of severe weather conditions such as volcanic eruption. The SL environment also provide students the opportunity to be innovative and create contents that can be transmitted freely within and out of the SL communities [16]. The ability of students to be innovative in SL is also echoed in the study of Chandra and Leenders [8]. However, it was deduced from the study that the innovative ability of students in SL, mimics that of their real-life ability to be innovative.

Using the SL environment for learning purposes can in some instances help save cost for educational institutions [1]. For example, if an institution cannot afford to build a physical laboratory for its students to conduct experiments, such experiment can be done in the SL environment for a much cheaper amount. The use of SL for learning purposes has, however, been subjected to mixed reviews. According to Olasoji and Henderson-Begg [1], issues relating to fairness in assessment and matching avatars with real students are some of the issues of contention with regards to the use of SL as an educational platform. Similarly, according to Fitch, Graves [17], the playful nature of SL is also a challenge, because it can result in students perceiving the activities that they do on SL as being playful rather than educational.

III. Theoretical Framework

The Technology Acceptance Model (TAM), developed by Davis [18] was adopted in this study. This model has been used by different authors to predict users' adoption and behavior towards technology [19]. TAM was adopted in this study because, in similar studies, it was also used. For example, in the study of Shen and Eder [20], which adopted a quantitative approach, TAM was used with some other constructs to examine the potential of VWs as a learning environment. Similarly, in the study of Chow, Herold [21], TAM was also the model adopted to describe the development and evaluation of SL for the learning of the method of acute airway management, known as rapid sequence intubation (RSI), in healthcare settings.

According to Wan Ismail, Chan [22], one of the main components of TAM that has been identified to influence the acceptance and use of technology is the Perceived Usefulness (PU) of the technology. As explained by the authors, PU is the extent to which an individual believes that the use of a particular technology would help improve their job performance. Therefore, to find out the extent to which students find SL as a useful learning platform, the PU and acceptance use constructs of TAM was used. This formed the theoretical construct to study the usefulness of SL as a learning platform, from the perspective of students.

IV. Methodology

This study used the exploratory design approach to explore and understand students' perceptions about the usefulness and challenges of adopting SL as a learning platform. A qualitative approach was also adopted in this study, hence semi-structured interviews were used as the means of data collection. The sample selection criteria was based on students' experience and interaction with SL, hence the purposive sampling technique was adopted. Data was collected from amongst the fourth year IS&T students at UKZN. This is because, one of the subjects offered by these students is taught using SL. There were 20 students in the class, however, saturation point was reached when 11 students were interviewed. At saturation point, no new information is acquired [23]. The interview instrument used consisted of 10 pre-set questions constructed in alignment with the literature and the model adopted in the study. The interviews allowed further probing questions to be asked so as to gain further insight to interviewee's responses. Interviews were conducted as confidential and were audio recorded and transcribed verbatim. The thematic data analysis approach was then used to identify themes from the interviews. The themes that emerged through the data analysis process were categorized into more focused but detailed categories that provided the needed insights that were relevant in achieving the study's set objectives.

V. Findings

The themes that emerged through the analysis process, provided the study's participants perception about the usefulness of SL as a learning platform. They also presented

some of the challenges faced in the use of SL as a learning platform. An additional theme that also emerged is the suggestion from the respondents as to how the challenges faced in the use of SL as a learning platform can be managed. These themes in relation to the study's objectives are presented in the next sections.

A. Perceived Usefulness of SL as a Learning Platform

The finding of this study shows that students find SL to be useful in their learning experience as it helped them engage better with their learning material which brought about a better understanding of the concepts that were being taught. Chow, Herold [21] did a study on the potential of Second Life to be used to enhance healthcare education. The study used TAM as a framework and the overall findings were that the participants perceived SL as a useful platform to enhance their learning. The findings of this study with regards to the usefulness of SL as a learning platform are presented below.

1) Facilitates Collaboration

Most of the study's participants were of the opinion that one of the usefulness of SL is that it facilitates collaboration. According to one of the participant, "*second life offers an engaging online experience in which collaboration amongst students is made possible*". Hansen, Richardson [24] and Jain, Iyengar [5] also shared this view in their study, in which they explained that VWs such as SL enhances real-time online collaboration and also allow users to design and construct environments that can be shared with others in a collaborative fashion. Being able to interact with lecturers and collaborate with classmates on different projects or assignments, is a feature considered by most of this study's participants, as making SL suitable for adoption as a learning platform. This is because the findings of this study shows that the use of SL helps students collaborate on their school tasks. This finding was particularly emphasized by one of the study's participant who explained that, "*within the SL platform, students were able to interact and collaborate with each other in order to understand the module's concepts, and also, in finding solutions to the assessment problems.* Similar to the participant's explanation is the discussion in the study of Olasoji and Henderson-Begg [1] which shows that SL can be a useful platform for conducting assessments.

The analysis of participants' responses also shows that the interaction, engagement and collaboration enabled through the SL platform, significantly facilitates students understanding of their study material. As explained by one of the participants, "*the use of SL helped improve our understanding of the study materials, and subsequently helped improve our performance in the module"*. This finding conforms to the suggestion in the study of Palomäki [10], which indicates that, VWs if adopted as a learning platform, can facilitate the provision of learning materials and virtual libraries to students. While at the same time, provide for students, the meeting place and environment to understand the learning materials.

2) Enables Learning Flexibility

One of the findings of this study is that SL provides some form of learning flexibility to students as it makes it possible for them to be able to complete their school tasks anytime during the day, and where resources (e.g. infrastructure) are available and adequate, from any location of their choice. This was also alluded to in the study of Hansen, Richardson [24], where it was stated that one of the benefits of VWs' adoption as an educational tool is that they can provide learning platforms that are available at any time and accessible from anywhere. Similarly, in the study of Olasoji and Henderson-Begg [1], it was stated that with SL, geographically dispersed students are provided the flexibility to learn in an environment similar to their traditional classrooms without forfeiting the ability to learn in their own time zone and at their own pace.

The flexibility offered by SL, according to Esteves, Fonseca [15], is a contributing factor as to why a number of educational institutions around the world are adopting SL as a learning tool. This flexibility, eliminates the physical boundaries of the conventional classroom [25]. Similarly, as identified by respondents in this study, it is also a form of remedy to the challenge identified in the study of Bosch-Sijtsema and Haapamäki [16], where it was stated that learning and knowledge development are often hindered by distance. Respondents in this study indicated that the flexibility offered by SL is changing the teaching and learning dynamics, especially in the area of distance learning. They also indicated that it is creating opportunities for universities to offer distance learning courses and programs to "non-resident" students. According to one of the participant, "*for students that want to attend to their school work, but are far away, SL can enable them do so, in the form of a distance learning course*". The ability of SL to facilitate distance learning is also agreed to in the study of Olasoji and Henderson-Begg [1] where it was stated that SL have the potential to transform educational processes, especially in distance education. Similarly, it was also agreed to in the study of Palomäki [10], where it was shown that SL can be used to support distance learning and interactive way of teaching.

3) Stimulates Creativity

One of the participant in this study explained that the use of SL in teaching is in itself a positive response to some of the changes required in education. This is because modern technology, its applications and platforms such as SL are enabling lecturers to adopt a more digital lecture room as compared to the traditional lecture rooms. These applications and platforms, as explained by another participants, are facilitating and improving students' creative abilities. Majority of the responses in this study pointed to the opinion that SL is an effective tool in improving students' creativity. According to one of the participants, "*the environment presented by SL enable creativity in learning and in finding solutions to assignment problems*". The creative opportunity afforded to students through the use of SL is also shown in the study of Bosch-Sijtsema and Haapamäki [16]. In the study, it was shown that one of the benefits of adopting any VW application and platform into education is that they help in enhancing the innovativeness of students.

The findings of this study also shows that with SL, students are able to extend their minds to find solutions. According to one of the participant "*SL offered me the freedom to stretch my mind, because in SL, I was the one constructing the environment that I was learning in and building what would allow me to learn better*". This finding is similar to what was stated in the study of Hansen, Richardson [24]. In the study, it was explained that students are required to actively build and interact in the SL environment. In the process of building, students are

afforded the opportunity to freely express themselves and stretch their minds as wide as they can. The discussion in the study of Hansen, Richardson [24] also shows that the adoption of SL as a learning platform presents an environments where creativity and learning among students is promoted.

B. *Challenges with the use of SL as a learning platform*

The findings of this study shows that SL have technical requirements that poses as a major setback to its adoption as a learning platform. In the study of Cheng [26], the technical requirements (such as high processing power, memory and powerful graphics card) of SL was also identified by educators to be a major drawback to its adoption. Similarly, Warburton [6] explained that the complexities and challenges of immersive environments such SL spans a range of technical and social intricacies that presents challenges to developers and educators trying to establish educational activities in the virtual space.

Being a web based platform, SL requires an adequate Internet connection and capacity, which majority of the respondents indicated was not always available. Majority of the respondents indicated that the Internet connection can sometimes be slow and thus slows down their work. A slow network connection can also cause technological glitches that could cause problems like delays which may negatively impacts the overall virtual world experience. According to one of the participants, "*the sometimes slow or inadequate Internet connection often results in lengthy waiting time when working in the SL environment. This consequently increases the actual productive learning time when working in the SL environment*".

Learning to use SL was also identified as a challenges. Most respondents indicated that learning to use SL can be time consuming as users need to fully understand the navigation controls needed to move around so as to complete tasks. The findings of this study also shows that there is always an assumption that students will be able to quickly learn and adapt to the use of SL, this however, is not always the case. Most respondents indicated that they struggled to use SL, and understand its controls, especially at the beginning of the module. This can be related to the study of Keelan, Ashley [27], in which they stated that, most of their study's participants had never before used the SL platform, and this was a challenge. For some participants in this study, the use of SL only became easier when they were already deep into the module and the semester. This, they believe affected their performance in the module. This findings thus indicate that there is the danger of over-estimating the technical capabilities of students. Similarly, according to Olasoji and Henderson-Begg [1], some researchers believe that a reasonable percentage of students may still be considered technologically challenged.

C. *Managing the challenges with the use of SL as a learning platform*

The findings of this study shows that, to manage the challenges faced in the use of SL as a learning platform, institutions need to provide adequate training and infrastructure that can support the adoption and use of SL as a learning platform. It can therefore be deduced that the success of SL as a learning platform is dependent on the resources and support available for its use [26]. Eschenbrenner, Nah [28] in their study also shared the same opinion. In the study, it was emphasized that the implementation of VWs such as SL should be supported by the right IT infrastructure, well trained staff and adequate management support. Respondents indicated that prior training sessions should be included as part of the teaching structure in the adoption of SL as a learning platform. This was also alluded to in the study of Keelan, Ashley [27]. It the study, support and training had to be provided to users in their recruitment into the use of the SL platform. They also indicated that the I.T infrastructure that is used to support SL should be well maintained and when necessary be reviewed and upgraded to align with the increasing number of students using the platform.

VI. CONCLUSION

SL have features that are highly interactive and that can enable the provision of dynamic feedback, learner experimentation and exploration. It is regarded as a versatile and useful educational platform through which learners can collaborate and engage with each other. It is also perceived as an efficient and useful tool in facilitating learning. In this study, students indicated that they prefer SL as a learning platform, as against most traditional way of learning. The use of avatars to represent and conceal users' real identity in the SL platform, was also found useful by this study's participants. Although, the use of SL comes with challenges such as its ease of use, navigation of controls and in some cases, inadequate IT infrastructure. The lack or inadequate training on how to use SL was also identified as a challenge to the use of SL as a learning platform. Overall though, the participants in this study indicated that SL is a good platform to be adopted in today's teaching and learning approach.

It is recommended that further studies should be conducted using a much wider population and a quantitative method applied to understand the extent of the impact of PU and its related factors on actual use of SL.

REFERENCES

[1] Olasoji R, Henderson-Begg S. Summative assessment in second life: A case study. Journal For Virtual Worlds Research. 2010;3(3).

[2] Upadhyay A, Rao N, Kornder J, Crawford C. Mechanism for Displaying Three-Dimensional Objects Alongside Video in Virtual Reality. 2017.

[3] Saeed N, Yang Y, Sinnappan S, editors. User acceptance of Second Life: An extended TAM with hedonic consumption behaviours. Proceedings of the 17th European Conference on Information Systems (ECIS'09); 2009.

[4] Loke S-K. How do virtual world experiences bring about learning? A critical review of theories. Australasian Journal of Educational Technology. 2015;31(1).

[5] Jain R, Iyengar S, Arora A, editors. Virtual world for education and entertainment. Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on; 2013: IEEE.

[6] Warburton S. Second Life in higher education: Assessing the potential for and the barriers to deploying virtual worlds in learning and teaching. British journal of educational technology. 2009;40(3):414-26.

[7] Dreher C, Reiners T, Dreher H, Dreher N. Virtual worlds as a context suited for information systems education: Discussion of pedagogical experience and curriculum design with reference to Second Life. Journal of Information Systems Education JISE. 2009;20(2):211-24.

[8] Chandra Y, Leenders MA. User innovation and entrepreneurship in the virtual world: A study of Second Life residents. Technovation. 2012;32(7-8):464-76.

[9] Inman C, Wright VH, Hartman JA. Use of Second Life in K-12 and higher education: A review of research. Journal of Interactive Online Learning. 2010;9(1):44-63.

[10] Palomäki E. Applying 3D virtual worlds to higher education. 2009.

[11] Chesney T, Chuah S-H, Hoffmann R. Virtual world experimentation: An exploratory study. Journal of Economic Behavior & Organization. 2009;72(1):618-35.

[12] Falloon G. Using avatars and virtual environments in learning: What do they have to offer? British Journal of Educational Technology. 2010;41(1):108-22.

[13] Dreher C, Reiners T, Dreher N, Dreher H. Virtual worlds as a context suited for information systems education: Discussion of pedagogical experience and curriculum design with reference to Second Life. Journal of Information Systems Education. 2009;20(2):211.

[14] Ali H, Ahmed AA, Tariq T, Safdar H, editors. Exploring Intentions to Use Second Life for Education at University of Bahrain. 2013 Fourth International Conference on e-Learning "Best Practices in Management, Design and Development of e-Courses: Standards of Excellence and Creativity"; 2013 7-9 May 2013.

[15] Esteves M, Fonseca B, Morgado L, Martins P. Improving teaching and learning of computer programming through the use of the Second Life virtual world. British Journal of Educational Technology. 2011;42(4):624-37.

[16] Bosch-Sijtsema PM, Haapamäki J. Perceived enablers of 3D virtual environments for virtual team learning and innovation. Computers in Human Behavior. 2014;37:395-401.

[17] Fitch TM, Graves MJ, Ferrell AC. Exchanging assets between on-line virtual worlds. Google Patents; 2008.

[18] Davis FD. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. MIS Quarterly. 1989;13(3):319-40.

[19] Amin H, Hamid MRA, Lada S, Anis Z. The adoption of mobile banking in Malaysia: The case of Bank Islam Malaysia Berhad (BIMB). International Journal of Business and Society. 2008;9(2):43.

[20] Shen J, Eder LB. Intentions to use virtual worlds for education. Journal of Information Systems Education. 2009;20(2):225.

[21] Chow M, Herold DK, Choo T-M, Chan K. Extending the technology acceptance model to explore the intention to use Second Life for enhancing healthcare education. Computers & Education. 2012;59(4):1136-44.

[22] Wan Ismail WK, Chan PHK, Buhari N, Muzaini A. Acceptance of Smartphone in Enhancing Patient-Caregivers Relationship. 2012. 2012;7(3):9.

[23] Fusch PI, Ness LR. Are We There Yet? Data Saturation in Qualitative Research. Qualitative Report, The. 2015;20(9):1408.

[24] Hansen S, Richardson J, Ellis A, editors. Virtual Worlds: Synopsis of user interfaces and accessibility initiatives. Ausweb08: Reflecting on the Past: Anticipating the Future: Proceedings of the 14th Australasian World Wide Web Conference, held in Ballina, NSW, 5-9 April, 2010; 2008.

[25] Singh N, Lee MJ. Exploring perceptions toward education in 3-D virtual environments: an introduction to “second life”. Journal of Teaching in Travel & Tourism. 2009;8(4):315-27.

[26] Cheng G. Exploring students' learning styles in relation to their acceptance and attitudes towards using Second Life in education: A case study in Hong Kong. Computers & Education. 2014;70:105-15.

[27] Keelan J, Ashley LB, Morra D, Busch V, Atkinson K, Wilson K. Using virtual worlds to conduct health-related research: Lessons from two pilot studies in Second Life. Health Policy and Technology. 2015;4(3):232-40.

[28] Eschenbrenner B, Nah FF-H, Siau K. 3-D virtual worlds in education: Applications, benefits, issues, and opportunities. Journal of Database Management (JDM). 2008;19(4):91-110.

M-Government Adoption Framework for Least Developed Countries: The Case of Malawi

John Edward Mtingwi
Department of Informatics
University of Pretoria
Pretoria, South Africa
johnmtingwi@gmail.com

Tendani Mawela
Department of Informatics
University of Pretoria
Pretoria, South Africa
mawelatendani@gmail.com

Awie Leornad
Department of Informatics
University of Pretoria
Pretoria, South Africa
awie.leonard@up.ac.za

***Abstract*—The Least Developed Countries (LDCs) are not producing and delivering services to stakeholders timely because of poor Information and communication technologies (ICTs) to support LDCs to handle important ICTs developmental challenges, such area as poverty reduction, financial inclusion, and illiteracy and health. There is greater need to support the accessibility of ICTs infrastructure in LDCs through the development of telecommunication infrastructures [94].**

The purpose of this research is to develop mobile government (M-government) Adoption Framework suitable for Least Developed Countries (LDCs) as most of the current technology adoption frameworks have been modelled around developed and developing countries. Literature has so far failed to provide any comprehensive technology adoption framework for LDCs' M-government adoption. Most stakeholders have no M-government experiences.

The failure necessitates the creation of a new M-government Adoption Framework that suits the LDCs to deploy M-government to leapfrog E-government. The proposed M-government Adoption Framework is intended for use in LDCs to deliver services through M-government. It will also assist decision makers to provide comprehensive and appropriate guiding principles to understand the challenges and opportunities of mobile government.

The research adopted a qualitative research using pragmatism philosophy, descriptive and explanatory approaches to enhance the presentation and understanding of the research findings. Case study research is employed, as it utilises realistic data from the proposed knowledgeable participants from selected organisations, ordinary citizens and focus groups in order to create appropriate contribution.

The most significance point is the availability of most factors that enable the adoption and implementation of M-government. Some of these factors include: mobile network providers, mobile phones, laptops and a minimal level of human capacity support to develop mobile content.

The proposed framework contributes new knowledge as it includes unexplored constructs and dimensions not existing in the current frameworks. The constructs of the proposed M-government frameworks would influence the generalisation of the developed framework for adoption and implementation of M-government in similar LDCs settings and others. The stakeholders have no M-government experiences.

Keywords—least developed countries, leap frogging, digital gap, pragmatism, e-government, m-government

I. INTRODUCTION

Least developed countries (LDCs) were formally established by the United Nations in 1971 [85]. These countries are countries with low-incomes. The low levels of incomes make LDCs to suffer from long term impediments to growth. LDCs have low human resource capacity development and are vulnerable both in social-economic and environmental stocks [85].

The telecommunication sector in the world has developed various electronic mobile devices which have penetrated the societies including LDCs. In general, a lot of individuals have mobile devices such as mobile cellular phones and laptops. These devices have enhanced people's communication methods since the last decade worldwide. These devices are used for communication amongst people of different income status. They are also linking citizens with government agencies and business entities.

As such, high power distance which is the extent to which the lower ranking individuals of a society are expected to accept the power unequally distributed [1]. It is principally utilised to study societal of inequalities of management between stakeholders in LDCs government organisation and the nation in general.

Illiteracy is unlikely expected to be successful to detail mobile government technology adoption because most people are using mobile phones because (Simple and affordable mobile phones are being used as a means to promote literacy for adults in Africa (Aker, 2009). Despite the fact that text messages are one-seventh the price of voice calls in Niger, the use of text messages has been relatively limited, in part due to high illiteracy rates). Towards the end of 2017, mobile-cellular subscriptions increased to about 700 million in LDCs with 70 per cent penetration. Over 80 per cent of the population in LDCs are living and covered with mobile cellular network [86]

The use of mobile devices as a channel to provide governmental services is known as mobile government. These devices range from tablets, PCs, laptops, cellular phones, and Personal Digital Assistants (PDAs). The idea of mobile government was initiated as a means to communicate with stakeholders [18] in developing and developed countries. LDCs have made continuous growth in terms of people having access to mobile ICTs.

The existing E-government literature highlights that E-government failures in least developed countries (LDCs) is due to lack of information communication technologies (ICTs) infrastructure [37]. E-government strives to improve participation of stakeholders in decision making. It makes public organisations to be accountable and transparent using electronic and mobile processing services (DESA 2018). It contributes to the production and delivery of public services.

As such, the services are more accessible, responsive and effective to stakeholders' requirements. Some the E-government models that did not work in LDCs are Layne and Lee (2001), Baum and DI Maio (2000), Hiller and Bélanger (2001), Wescott (2001) and Gartner's Four Phases Model.

Reducing and bridging the digital divide is identified as a critical factor for improvements in E-government implementation. The average world E-government Development index (EGDI) improved from 0.47 in 2014 to 0.55 in 2018 because of continuous improvement of indices of subcomponents. This suggests that globally, there has been steady progress in improving E-government and public services online provision. But despite some development gains and major investments made in several countries, the e-government and digital divides still persist. Among the Fourteen countries in the Low-EGDI group are African and belong to the least developed countries.

The average E-government Development index (EGDI) for LDCs increased a bit from 0.21 in 2014 to 0.24 in 2016. The number of LDCs with EGDI values of less than 0.25 remained constant at 29 in 2016. Most of these LDCs have extremely low ratings for telecommunications infrastructure.

UN-OHRLLS [37] indicated that LDCs scored lowest average of 0.24 in 2016 at the E-government Development Index (EGDI) compared to lower middle-income countries with an average of 0.41 of EGDI [36]. The lowest score is owing to inadequate ICTs infrastructures. LDCs are not making significant progress to realise advanced E-government development phases which foster the delivery of e-services, provision of government information and e-participation, despite having in place some elementary E-government structures [37]. LDCs have made continuous growth in terms of people having access to mobile ICTs.

II. Problem Statement

Governments in LDCs are not delivering quality services to stakeholders on time. The failure has come in due to insufficient finances, inadequate skilled manpower, lack of appropriate devices for connectivity and lack of private-public partnerships to develop national telecommunication network backbones [26].

The absence of a well-built national telecommunications network backbone results in poor information technology communication infrastructure. LDCs are ranked the lowest according to the world E-government ranking [36], [37]. Despite ranked lowest on the world E-government development index, governments in LDCs started E-government implementation procedures to improve service delivery to various constituents.

To achieve the aims of E-government in Malawi (Case Study), since ICT infrastructures are not developed, another method is supposed to be explored to support E-government to deliver government services to stakeholders. One method is the adoption and use of M-government. The use of mobile phones and other portable communicative devices [23] to distribute services and information is very possible taking advantage of the existing mobile network providers [22].

The primary problem is why LDCs have no M-government Adoption Frameworks that suit them as the existing frameworks have failed to adopt M-government implementation.

Qualitative methodology clearly shows that the existing models have limitations, e.g. the use of Roger's [88] diffusion of innovation model which uses a quantitative methodology. Quantitative study effectively confirmed that Roger's innovation characteristics obstructed adoption decisions Richardson's [86].

The models' characteristics frequently overlap. They are discoursed as strictly aligned characteristics that are regularly difficult to differentiate. Employing qualitative analysis of the transcribed data certified a richer and deeper appreciation of how these inventive features affect the selections to adopt based on challenges and barriers experienced by the end user

Lack of proper M-government adoption frameworks is perpetuating the digital gap [24], [26]. As a result, quality differences of available fixed and mobile networks and the basic connectivity, digital gaps are prevalent between and within LDCs and other countries [26].

The digital gap problem between stakeholders and governments in this research is referred to as: Stakeholders and government information sharing gap [29]. Information sharing gap, refers to governments' failure to effectively share information and other services with the relevant stakeholders on time because of variances in ICTs development, in LDCs, as well as other countries [24], [26].

The digital gap and exclusion in the utilisation of the innovative mobile technologies is due to different factors including: cultural, economic, differences in geographical features, perceptions, age and gender gaps. In any case Internet connectivity remains a costly method of obtaining and exchanging information in LDCs. But the mobile and televisions ubiquitous can consequently contribute to the reduction of digital gap.

The research further sought to develop LDCs' specific proposed M-government Adoption Framework that would assist them to adopt mobile government. Information sharing between government and stakeholders through utilisation of mobile technologies could improve if mobile government is adopted.

III. Purpose of Research

The purpose of the research is to come up with M-government Adoption Framework for adoption of M-government in LDCs. Literature has failed to provide comprehensive Technology Adoption framework for LDCs' M-government deployment [1]. For instance, the current models such as Diffusion of Innovations (DIT), Technology Acceptance Model (TAM), and Technology Task Fit (TTF) have not included some important constructs or variables necessary for M-government Adoption Framework in LDCs.

Some of the variables missing in the models include ICT technical expertise, organisational compatibility, mobile technology usability, critical environmental factors such as (pricing of mobile items, security, legal related issues and education level) and ICT project development plans. Economic Variables (ICTs Expenditure), Gross Domestic Product (GDP) per capita, Unemployment Rate and Annual GDP growth.

Human Development Index in Fig 2 illustrates the constructs that will assist to address the gaps. Information seeking behaviours in LDCs indicates formal to informal methods of gathering of information readily available and stakeholders centered services [90]. Knowing the numerous challenges LDCs are facing, stakeholders are changing from formal to informal communication strategies [91].

The great interest of stakeholders is learning the available mobile technologies, mobile information technology and mobile access technologies. The accessibility of information on the Internet is making the world wide open, access to databases and academic through mobile devices [92]. Information Seeking is using mobile technologies and Web 2.0 applications on the wider social contexts [93].

The gaps found in these models for adopting and implementing M-government enhance the need to develop a new M-government Adoption Framework that would suit the LDCs to deploy M-government. The framework's constructs have relevant propositions which identify the factors that would influence the implementation of M-government. The proposed M-government Adoption Framework of this research is intended for use in LDCs that have interest in delivering services through M-government. The adoption framework will contribute new knowledge because the framework includes unexplored constructs and dimensions missing in the existing frameworks. The results of the research will assist decision makers who require research outcomes that provide comprehensive and appropriate guiding principles to understand the challenges and opportunities.

IV. Research Motivation and Significance

E-government offers prospects to use information and communication technologies to support government process administration of government services, from government institutions and deliver the services to stakeholders. The factors which enabled the researcher to embark on the research study in the case study country (Malawi) to develop M-government adoption framework are:

- The presence of mobile network providers, mobile phones and other portable devices for communication in the case study.
- Presence and utilisation of mobile applications in private sector where government organisations can learn from e.g. banking sector, international nongovernmental organisations in health sector.
- Political leadership emphasis that the people should use mobile technologies to participate in formulation of policies and decision making at various levels of government structures in democratic political system,.
- The existing M-government technology adoption theories or models have not been used in least developed countries (LDCs) landscape and are not consistent, and there is conspicuous lack of mobile government literature for LDCs.

The research is significant because the outcomes would guide government's political and technical leadership to make informed decisions. The outcomes will help academic and industrial communities in information systems to appreciate the need for M-government in LDCs to enhance communication and service delivery amongst stakeholders.

V. Research Contributions

The practical implication is that the proposed M-government adoption framework would assist in implementation of mobile government in LDCs. The implementation of the framework would help identify subject elements or components relevant to the stakeholders. The knowledge gained from this research enhances industry practice in the deployment of mobile network providers and development of contents to suit the needs of stakeholders. The knowledge gained supports policy and decision-making bodies to appreciate the use of mobile technologies towards practical adoption of M-government adoption in LDCs.

The outcomes may also enhance the adequate understanding of government technical staff of what M-government is. These technical staff will be involved in the implementation of M-government and they will help to transfer M-government knowledge to others. The outcomes will assist the public to appreciate the importance of using electronic mobile devices in government organisational activities. The understanding of the outcomes may also help to change the mindset, which is officials' traditional way of providing services to relevant stakeholders. This change will reduce resistance to adopt the use of mobile technologies and adoption of M-government in LDCs' governments.

VI. Mode of Inquiry and Research Approach

The research utilised qualitative researching approach and a pragmatism philosophical paradigm to source the required data. The primarily focus was on extracting data from selected participants and existing literature. The discovered data were linked to their theoretical model concepts or constructs for validation or verification. The theoretical constructs would not be measured statistically. The research's methodology is as follows:

i. Malawi is a research setting (case study) because the country belongs to African least developed countries [5], [27], [37], [28]. Malawi was preferred in the evaluation of the constructs or variables which would be influential to M-government adoption in African LDCs.
ii. Pragmatic philosophical approach is suitable for conducting the research, because worldviews are believed to be constructed socially on a shared agreement [72].

The principal idea of qualitative research is the discovery of the meanings rather than quantifying. The case study used, is compatible with [4] qualitative research approach. Qualitative research method was adopted for the direct interactive interviews to collect data from participants and reviewing useful literatures [72], [33]. Qualitative research focuses on exploring questions such as what, how and why.

The developed theoretic framework is a guiding theoretic lens of a researcher to develop or formulate the investigative design of the research for data gathering, transcribing, cleaning, sorting and analysing [40], [43], [39], [72].

The pragmatist paradigm is supported because it provides for investigation of actions and analysis of research outcomes to improve social reality. Pragmatic researchers deduce the role of prior issues within a specific societal context [6]. It aims at intervening the worldview through actions of human beings [7]. Pragmatism argues that the most important determinant of the research philosophy adopted is the research question. It is a practical approach and integrates different perspectives to collect and interpret the data [3]. Pragmatism employs qualitative strategy because the phenomenon requires an in-depth investigation, explanation and description [2], [4]. Pragmatism appreciates the context in which actions can be undertaken after knowledge has been created to solve existing problems [6], [8]. Figure 1 depicts level of mobile population in LDCs.

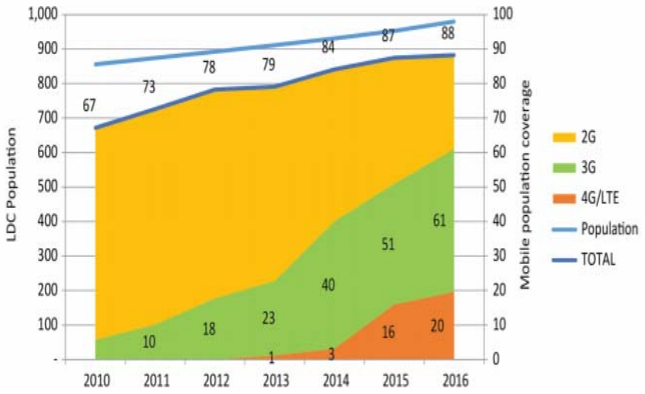


Figure 1: Mobile population coverage in LDCs (Source: [94])

VII. Malawi Case Study Context

Internationally Malawi is categorised a LDCs because of the low-income of 1.61 ICT Development Index (IDI) value (ITU 2015a; ITU 2016a), low human capital and economic vulnerability. Figure 2 illustrates low quartiles of IDI values and that many landlocked countries (LCCs) and LDCs are African countries with an average of 2.48 IDI values (ITU 2015a; UNCTAD 2015; ITU 2016a). See the countries marked with yellow colour. Malawi is additionally a landlocked country (a country without sea ports for exportation and importation) and this contributes to the fragile economy, low individual incomes and low human skills.

The fragile economy incapacitates the peoples' buying power of mobile technologies in LDCs. Telecommunication infrastructures in Malawi are not adequate and provide 1.07 fixed telephone lines for every 100 persons (United Nations 2012).

Malawi's E-government is struggling to take off because it requires huge financial infrastructure investment. Malawi government has not adopted M-Government but few private and international organisations have adopted mobile applications such as in health sector for mobile information processing in hospitals, agriculture sector for agribusiness information to farmers and other stakeholders and banking sector for banking services. At national level, ICTs development has witnessed the growth of two main mobile network providers namely TNM and Airtel companies.

The status of E-government is that it has many integrated systems that provide solutions for a particular organisation. The anticipated benefits of Malawi E-government were to provide stakeholders with services in time through effective and efficient new ways and socio-economic development. Some challenges of Malawi E-government are financial problems, lack of political and technical leadership, difficult terrains and luck of highly ICT skilled professionals.

Malawian government organisations need to adopt M-government because it is anticipated that M-government would assist government to provide services to the research question. This will be achieved through:

i. Defining and naming themes to develop detailed analysis of each theme and choosing informative name for each theme.
ii. Putting together through writing the analytic description and data extracts. Finally contextualising existing literature in relation to the analysed transcribed data.

VIII. LDCs Leapfrogging E-Government

In LDCs, E-government has disappointing challenges [47], [71] that have forced people to migrate mobile technological devices, particularly cellular phones to have access to basic services as they have lower costs compared to computer costs. MTs are aiming at the minimisation of societal gap for information access which fixed phones failed to bridge. The digital gap can be bridged through the use of wireless or mobile technologies. The MTs can create M-government to provide ubiquitous government.

MTs are driving rapid ICTs network infrastructure development, affordability and skills which are essential for LDCs that want to achieve [94]. ITU [94] is demonstrating that numerous LDCs are providing impressive leapfrogging. This is triggering progressive development of policies. Leapfrogging in LDCs is boosting mobile broadband coverage of the population at over 90 per cent from 10 per cent more than seven years ago, see figure 2. What they both have in common are policies, for enabling leap-frogging. Many LDCs provide remarkable leap-frogging to assist them producing and delivering services in remarkable levels to stakeholders. What they both have in common are policies enabling leap-frogging.

The calls to access universal and affordable the Internet in LDCs, is suggesting that the main are obstacles to ICT infrastructure developments are high costs enable LDCs to be connected. As the expansion of the broadband connectivity continues the prices will be dropping. There is ability and desire to Internet utilisation although finances seem to be critical barriers.

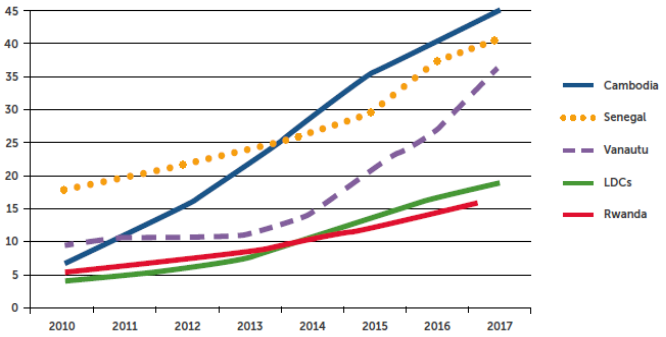


Figure 2: broadband speeds enhancing Leap-frogging

MTs are prerequisites for LDCs to leapfrog E-government [36], [74]. Leapfrogging is capable of bypassing traditional E-government developmental procedures to cut costs and timeframe [36]; [19]. Leapfrogging would preserve all crucial requirements to improve communication in LDCs and that leapfrogging to M-government without depending on E-government backbones [22].

Government should be enthusiastic to engage MTs to deliver the required services to stakeholders. M-government structure is built on E-government structure where it exists or M-government can be developed from the scratch where E-government is non-existent, for example in rural sites where there is no any physical ICT infrastructure to support connectivity of ICTs.

IX. Theoretical Framework

The theories such as the technology acceptance model (TAM) [77], [78], [79], The unified theory of acceptance and use of technology (UTAUT) [42], Task-technology fit (TTF) [44],[80], Diffusion of Innovation theory (DIT) [76] are the widely used technology adoption and acceptance theories from the existing reviewed literature. However, the models are not addressing the LDCs needs holistically. Therefore, there is need to create a proposed M-government Adoption Model that fits LDCs to address LDCs governments failure of delivering quality services to stakeholders efficiently. The failure is due to inadequate skilled manpower, insufficient finances and lack of appropriate connectivity devices. There are no comprehensive private-public partnerships to develop national telecommunication network backbones [81], [42], [82], [83]. For instance, TAM focuses on perceived ease of use and perceived usefulness of technology but does not include important external moderating variables or factors [41], [42], [50] that result into meaningful perceived ease of use and perceived usefulness of technology that would make M-government usable. UTAUT's major contribution is the integration of social and cultural factors.

Kim and Crowston [49] identified some of the limitations of ICT adoption theories as follows: ICT adoption and use theories were considered as one-off event disregarding the evolving ICT use dynamics (users' behaviour on continuous or discontinuous use of ICTs) after adoption, use of the similar or same constructs in different research designs are inconsistent [51] and ICT adoption studies fail to resolve the context understudy because they only focused on organisational contextualisation. The limitations of these popular preceding ICT adoption models inspired the formulation of the proposed M-government adoption framework for LDCs. In the proposed adoption framework, tasks or services, demographic pointers and critical factors are indicated which directly or indirectly affect the deployment of M-government.

The proposed M-government Adoption Framework as shown in Figure 3 is consisting of constructs from M-government Value Chain Model [84], TTF, TOE, TAM and UTAUT, M-government framework [16] and the stakeholders in the framework imply that the M-government if adopted would gradually be implemented to cover all national levels regardless of peoples' vicinity and responsibilities.

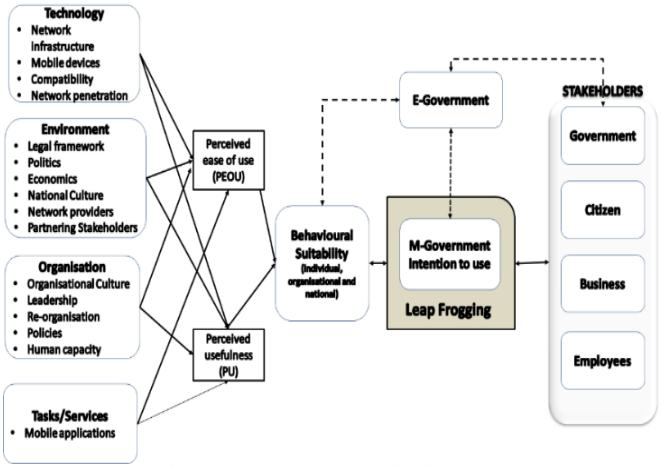


Figure 3: Proposed LDCs M-government Adoption Framework

X. The Proposed M-Government Adoption Framework Thematic Summaries for LDCs

The proposed M-government Adoption Framework is grouped into four major components or dimensions as outlined: Technology, Environment, Organisation and Tasks/Services which feed into PEOU and PU for validation to determine their importance for their behavioural suitability (at individuals, organisational and national levels) for the intention to develop and utilise M-government. The summary of components are discussed below:

General aspects: Most of the research participants have the knowledge of what M-government is but have not used it or do not have practical experience as it is not available in the case study country. Some have used the mobile applications private organisations such as banks and mobile network providers. The brief is that if the stakeholders in the society are aware of the benefits of M-government, they would influence its adoption. The ages of individuals and literacy levels would influence the adoption and spread of mobile services. For instance, the youths are very active and versatile with mobile technologies compared to older generations. The stakeholders would find M-government useful because they will be interacting with government on real-time basis. It is believed that the M-government will enhance the implementation of E-government in LDCs

Technology Adoption component: This component indicates the mobile network infrastructure development. This includes the approximate number of the population that owns the mobile devices, compatibility of different mobile networks and their related application systems and the network penetration rate in the country [45], [44]. The component also consists of compatible MTs, mobile related software, network infrastructure which is wireless in nature and secure environment [52] and private information storage and transmission [53], [54]. Interoperable and dependable technologies influence M-government adoption [35], [53]. Mobile tools' connectivity encourages the adoption of M-government in many countries [17], [55] and emerging MTs' innovative standards are used to evaluate M-government adoption.

Environment component: This component explores legal framework for mobile technology use in the country [56], [45], political setup that can support the implementation of M-government, the economic empowerment of governments and individuals to afford the acquisition and usage of mobile devices [57], [35]). It explores how national culture will affect the deployment of M-government. Finally to explore the reliability of mobile network providers in LDCs and to determine key elements that will enable stakeholders to participate in M-government. Governments in LDCs should deliberately provide laws that promote use of mobile tools [58]. Legal tools regulate the utilisation of MTs to disseminate different informational issues [59] at a very short period [60]. Partnering with other players in the MTs world will greatly enhance M-government adoption in LDCs [61].

Organisation component: Consists of government organisational culture towards M-government [63], [62] government leadership style towards the use of mobile technologies in offices [35]). It will also deal with the expected re-organisation of government organisation setup in line with the existing and upcoming policies such as E-government national framework to suit mobile government applicability. The E-government national framework should have the support of knowledgeable and committed leadership at all applicable levels [64]. It will also look at the human capabilities to manage the whole process of M-government implementation [45]. Users of the organisational information need trusted new mobile technologies which preserve their integrity [65]. It is advisable for LDCs to start developing small M-government projects because of their resource constraints and this approach will enable government workers to smoothly understand and adapt the new mobile systems and organisational restructuring [34]. The smooth understanding of M-government development in LDCs will help the key government workers to educate other government workers and stakeholders to adopt M-government [66].

Tasks or Services: The component will explore the likely candidates of mobile applications that will be implemented on M-government in line with government organisational foundations. This requires migrating from existing e-services to m-services or transforming manual systems straight to mobile applications [44]. Governments need to identify tasks that fit into the mobile technologies, organisational needs and the entire environment to provide positive and productive performance for stakeholders [67]. Some of the likely mobile services which will run on M-government infrastructure include the following: m-health, m-disaster, m-participation, m-education, m-agriculture, m-security, m-voting, and m-business [68], [21], [69], [70].

XI. RESEARCH LIMITATIONS AND FUTURE RESEARCH

The major limitations of the research is that no one has M-government practical experiences in the country and that there is no sufficient research of mobile government adoption for LDCs. The proposed framework provides new lenses through which researchers can study M-government adoption and implementation in LDCs. The participants worry of their privacy and security of personal data requires further research on cyberspace security in line with least developed countries.

XII. CONCLUSION

Least developed countries governments are utilising mobile technologies to improve information communication through social network in response to the needs of the stakeholders. LDCs governments should use whatever current mobile technologies they have to process the information and deliver services to the beneficiaries who own mobile devices. The mobile devices can offer m-governance platforms to bridge stakeholders and governments in real-time environment. The use of mobile devices enable governments in LDCs to make huge investment savings.

Mobile network providers could work on content in collaboration with all relevant stakeholders and provide direction on appropriate contents which suit mobile devices for the betterment of mobile networks connectivity. Government deliberate policy plans should be developed to develop capacity of ICT professionals in LDCs governments in areas of mobile applications development and application of privacy and security issues.

Both professional and political leadership should have the willingness to utilise ICTs in government process to provide services for social-economic development. The Ministry of ICT through Department of E-government should champion adoption of M-government in various public agencies.

The legal frameworks are important in LDCs to regulate the mobile technology telecommunication industry such as social works, internet cyberspace activities. For instance, the case study country (Malawi) has Communication Act 2016, Electronic Transaction and Cybersecurity Act 2016, National ICT policy but there is no serious enforcement or policing of these legal frameworks to benefit the stakeholders.

Economically, the participants complain about mobile network providers' high pricing of mobile phones' services. At the same time mobile network providers complain that government through the regulator is demanding or charging huge various tax fees which are passed on to mobile phones user. These economic stakeholders' complaints need government interventions of sustainable economic policies that lessen mobile device users' financial burden and promote mobile technological investment.

Governments in LDCs should involve the stakeholders in the selection of mobile applications and their contents for processing and transmission. For Malawi, interaction challenges between government organisations and their stakeholders could be improved if M-government could be adopted and operationalised. It should be appreciated that M-government is a viable option for the transformational E-government advances in Malawi.

LDCs governments should learn from private organisations which are using mobile applications in the country or from other countries. The study has shown that Malawi has all the needed factors (mobile technology establishments) for the adoption of M-government using the proposed M-government Adoption Framework. The stakeholders are enthusiastic to engage government services

using mobile devices and that government should lead by supporting and promoting mobile ICTs usage. The numbers of mobile subscribers have increased approximately in million that the government of Malawi could engage to pilot the appropriateness of any mobile application on the adopted M- government.

All research study participants' responses highlighted same worldwide challenges of M-government as indicated in [75], [53], [16], [23]. In many LDCs, these challenges can be addressed along the operational M-government processes if adopted.

REFERENCES

[1] Munyoka, W., and Manzira, M. F. "From E-Government to M-Government - Challenges Faced by Sub-Saharan Africa. Dubai, UAE, Proceeding of the International Conference on Computing Technology and Information Management". 2014.

[2] Robson, C., "Real World Research. 2nd ed. Oxford: Blackwell". 2002.

[3] Saunders, M., Lewis, P., and Thornhill, A. "Research Methods for Business Students. 5 ed. Essex CM20 2JE, England: Pearson Education Limited. ". 2009

[4] Yin, R. K. ´Case Study Research: Design and Methods. 3rd ed. Newbury Park. CA: SAGE." 2003.

[5] Creswell, J. W. "Research Design: Qualitative, Quantative, and Mixed Methods Approaches. 3rd ed. Thousands Oaks, California: SAGE." 2009.

[6] Goldkuhl, G. 2004. Meanings of Pragmatism: Ways to conduct information systems research. Sweden, 17-18 March, Linköping University, pp. 1-15.

[7] Braa, K. and Vidgen, R. "Interpretation, Intervention and Reduction in the Organisational Laboratory: A Framework for in-context Information Systems Research. Accounting, Managaement and Information Technology", 1999. Volume 9, pp. 25-47.

[8] Goldkuhl, G. "Pragmatism vs Interpretivism in Qualitative Information Systems Research. European Journal of Information Systems, 21(2), ", 2012. pp. 135-146.

[9] Cavana, R. Y., Delahaye, B. L., and Sekaran, U. "Applied Business Research: Qualitative and Quantitative Methods. Milton, Queensland, Australia: John Wiley and Sons." 2001.

[10] Sekaran, U. "Research Methods for Business, a Skill Building Approach. 2 Ed. New York: John Wiley and Sons." 2003.

[11] Bryman, A. "Social Research Methods. 4th Ed. New York: Oxford University Press Inc." 2012.

[12] DeCuir-Gunby, J. T., Marshall, P. L., and Mcculloh, A. W. "Developing and Using a Codebook for Analysis of Interview Data: An Example from a Professional Development Research Project. Field Methods, 23(2), 2011. pp. 136-155.

[13] Braun, V., and Clarke, V., "Using Thematic Analysis in Psychology. Qualitative Research in Psychology, 3(2)", 2006.pp. 77-101.

[14] Attride-Sterling, J. "Thematic Networks: An Analytic Tool for Qualitative Research. Qualitative Research, 1(3)," 2001. pp. 385-405.

[15] Carroll, J. "Risky Business: Will Citizens accept M-government in the Long Term. Mobile Government Consortium International LLC. Brighton, UK: University of Sussex" 2005.

[16] Carroll, J. "What's in It for Me?': Taking M-Government to the People. Slovenia, 19th Bled eConference on eValues, June, . 2006. pp 5 -7.

[17] Chen, L. and Nath, R. "A Framework for Mobile Business Applications. International Journal of Mobile Communications, 2(4)." 2004, pp. 368-381.

[18] Gosebo, N. and Obono, S. D. E. "Revisiting the Theory of Planned Behaviour for the Preparation of the Adoption of Municipal e-services in Less Developed Countries. SAICSIT, "2012. pp. 159-166.

[19] ILO, "World Employment Report 2001-Key Findings, Geneva, and March: International Labour Organisation." 2001.

[20] Informata, "Mobilising Public Services in Africa: The M-government Challenge. Informa UK Ltd. All rights reserved," 15 September, 2011. pp. 30-40.

[21] Lallana, E.,"E-Government for Development M-Government: Mobile/Wireless Applications in Government." 2004

[22] Mtingwi, J. and Van Belle, J. P. ´The State of E-Government and M-Government Readiness in Malawi. International Journal of Information Technology and Computer Science (IJITCS)", 2012. 6(1), pp. 1-20.

[23] Kushchu, I. and Kuscu, H. "From E-government to M-government: Facing the Inevitable. Trinity College, In Proceeding of European Conference of E-government (ECEG)". 2003.

[24] ITU, "Measuring the Information Society (MIS) Report, CH-1211 Geneva Switzerland: International Telecommunication Union." 2015a.

[25] ITU, "ICT Facts and Figures, Geneva, Switzerland: International Telecommunication Union" 2015.

[26] ITU, "Measuring the Information Society Report, Geneva Switzerland: International Telecommunication Union Place des Nations CH-1211". 2016a.

[27] ITU, "ICT Facts and Figures 2016, Geneva 20 - Switzerland: International Telecommunication Union Place des Nations 1211." 2016.

[28] UNCTAD, "The Least Developed Countries Report 2016. UNCTAD secretariat at: Palais des Nations, CH-1211 Geneva 10, Switzerland. United Nations Publication. New York." 2016.

[29] Khan, M. A. "Exploring the Push and Pull Drivers in M-Government Framework that Influence Acceptance of Services on Mobile Devices. International Journal of Computer Science and Network Security (IJCSNS), 16(2)," 2016. pp. 23-26.

[30] Moon, J. "From E-government to M-government? Emerging Practices in the Use of MT by State Governments". 2004.

[31] Markets and Research, "Malawi - Telecoms, Mobile and Broadband - Statistics and Analyses 2016 - Research and Markets" 2016.

[32] MACRA, "Country report on telecommunication status in Malawi, Blanytre, Malawi: Malawi Communication Regulatoty Authority". 2012.

[33] Myers, M. D. "Qualitative Research in Business and Management. London: SAGE." 2013.

[34] Syamsuddin, I. "Evaluation of E-government Initiatives in developing countries: An ITPOSMO Approach. International Research Journal of Applied and Basic Sciences," 2011. 2(12), pp. 439-446.

[35] Tornatzky, L. G., and Fleischer, M. 1990. The Processes of Technological Innovation. Massachusetts: Lexington Books, Lexington.

[36] UNDESA, "E-government for the future we want: United Nations E-government Survey 2014, New York: United Nations Department of Economic and Socila Affairs". 2014.

[37] UN-OHRLLS, "State of the Least Developed Countries, New York, NY 10017, USA: United Nations Office of the High Representative for the Least Developed Countries, Landlocked DevelopingCountries and Small Island Developing States". 2016.

[38] Varshney, U., and Vetter, R. "Emerging Mobile and Wireless Networks. Communications of the ACM," 2000. 43(6), pp. 73-81.

[39] Walsham, G. Interpretive Case Studies in IS Research: Nature and Method. European Journal of Information Systems," 1995. 4(2), pp. 74-81.

[40] Walsham, G. "Doing Interpretive Research. European Journal of Information Systems." 2006. Volume 15, pp. 20-30.

[41] Sun, H., and Zhang, P. "The Role of Moderating Factors in User Technology Acceptance. International Journal of Human-Computer Studies," 2006. 64(2), pp. 53-78.

[42] Venkatesh, V., Morris, M. G., Davis, G. B., and Davis, F. D. "User Acceptance of Information of Technology: Toward A Unified View. MIS Quarterly", 2003. 27(3), pp. 425-478.

[43] Miles, M. B., and Huberman, A. M. "Qualitative Data Analysis: An Expanded Source book. Thousand Oaks: SAGE." 1994.

[44] Gebauer, J., and Shaw, M. "A Theory of Task/Technology Fit for Mobile Applications to Support Organisational Processes. Working Paper,. Illinos: University of Illinois at Urbana- Champaign". 2002.

[45] Fasanghari, M., Amalnick, M. S., and Khatibi, V. "A Proposed Framework for M-government Development Assessment. s.l., Networked Computing (INC)", 2010. pp. 1-5.

[46] Banerjee, P., and Chau, P. Y. K. "An Evaluative Framework for Analysing E-government Convergence Capability in Developing Countries. Electronic Government, an International Journal," 2004. 1(1), pp. 29-48.

[47] Ndou, V. "E–Government for Developing Countries: Opportunities and Challenges. The Electronic Journal on Information Systems in Developing Countries", 2004. 18(1), pp. 1-24.

[48] Norris, D. F., and Moon, M. J. "Advancing E-government at the Grassroots: Tortoise or Hare? Public Administration Review", 2005. 65(1), pp. 64-75.

[49] Kim, Y., and Crowston, K. "Technology Adoption and Use Theory Review for Studying Scientists' Continued Use of Cyber-

infrastructure. New Orleans, LA, Proceedings of the American Society for Information Science and Technology (ASIST) Annual Conference, October 2011". 2011.
[50] Chin, W. W., Marcolin, B. L., and Newsted, P. R. "A Partial Least Squares Latent Variable Modeling Approach for Measuring Interaction Effects: Results from a Monte Carlo Simulation Study and Voice mail Emotion/Adoption Study. Information Systems Research 2003, 2003. 14(2), pp. 21-41.
[51] Limayem, M., Cheung, C. M., and Chan, G .W. W. "Explaining Information Systems Adoption and Post-Adoption: Toward an Integrative Model. s.l., International Conference on Information Systems" 2003. (ICIS 2003 Proceedings).
[52] Shah, J. R., and Murtaza, M. B. "Effective Customer Relationship Management through Web Services. 2005. pp. 98-109.
[53] Sheng, H., and Trimi, S. "M-government: Technologies, Applications and Challenges. Electronic Government, an International Journal, 2008. 5(1), pp. 1-18.
[54] Kuan, K. K. Y., and Chau, P. Y. K. "A Perception-Based Model for EDI Adoption in Small Business Using a Technology-Organisation-Environment Framework. 2001. 507-512.
[55] Malladi, R., and Agrawal, D. P "Current and Future Applications of Mobile and Wireless Networks. Communication of the ACM", 2002. 45(10), p. 144–146.
[56] Delmas, M. A. "The Diffusion of Environmental Management Standards in Europe and in the United States: An Institutional Perspective. Policy Sciences", 2002. 35(1), pp. 91-119.
[57] Crook, C. W., and Kumar, R. L. "Electronic Data Interchange: A Multi-Industry Investigation Using Grounded Theory. Information and Management", 1998. 34(2), pp. 75-89.
[58] Xu, S., Zhu, K., and Gibbs, J." Global Technology, Local Adoption: A Cross-Country Investigation of Internet Adoption by Companies in the United States and China. Electronic Markets", 2004. 14(1), pp. 13-24.
[59] Vincent, J., and Harris, L. "Effective use of mobile communications in eGovernment: How do we reach the tipping point?. Information, Communication and Society", 2008. 11(3), pp. 395-413.
[60] Hermanns, H. "Mobile Democracy: Mobile Phones as Democratic Tools. Politics", 2008. 28(2), pp. 74-82.
[61] Zhu, K., Kraemer, K. L., and Xu, S. ´E-Business Adoption by European Firms: A Cross-Country Assessment of the Facilitators and Inhibitors. European Journal of Information Systems", 2003. 12(4), pp. 251-268.
[62] Hofstede, G. "Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organisations across Nations. 2 ed. Thousands, California: SAGE". 2001.
[63] Hofstede, G. "Cultural Dimensions: South Africa Cultural Dimensions". 2009
[64] Krishna, S., and Walsham, G. ´Implementing Public Information Systems in developing countries: learning from a success story. Information Technology for Development", 2005. 11(2), pp. 123-140.
[65] Rao, S., and Troshani, I. "A Conceptual Framework and Propositions for the Acceptance of Mobile Services. Journal of Theoretical and Applied Electronic Commerce Research", 2007. 2(2), pp. 61-73.
[66] Khalifa G., Irani Z., Baldwin L. P., and Jones S. "Evaluating Information Technology with You in Mind. The Electronic Journal of Information Systems Evaluation", 2004. 4(5), pp. 246-252.
[67] Irick, M. L "Task-Technology Fit and Information Systems Effectiveness. Journal of Knowledge Management Practice". 2008, 9(3), pp. 1-5.
[68] Rain, R., and Maarja, S. "Mobile Services in Tartu Existing services, best practices methodology and suggestions for future development, Tartu, Estonia: Mobile Solutions Ltd. 2005.
[69] Ghyasi, A. F., and Kushchu, I. "M-government: Cases of Developing Countries. 2004. [Online] Available at: http://unpan1.un.org/intradoc/groups/public/documents/apcity/unpan040050.pdf [Accessed 12 March 2015].
[70] Ewan, M. "Dubai has ambitious targets for m-government system". 2006.
[71] Heeks, R. "Most eGovernment-for-Development Projects Fail: How Can Risks be reduced? (IGovernment working paper series; No. 14). Manchester, University of Manchester", 2003. pp. 1-17.
[72] Creswell, J. W. "Research Design: Qualitative, Quantative, and Mixed Methods Approaches. 3rd ed. Thousands Oaks, California: SAGE." 2009.
[73] Global Economy, "The Global Economy". 2016. [Online] Available at: http://www.theglobaleconomy.com/Malawi/Mobile_phone_subscribers/ [Accessed 16 October 2016].
[74] El-Kiki, T., and Lawrence, E. "Government as a Mobile Enterprise: Real-time, Ubiquitous Government. Las Vegas, NV, 2006 Proceedings. Third International Conference on Information Technology (ITNG): New Generation 2006," 2006. pp. 320-327.
[75] Lanwin, B. A Project of Info Dev and The Center for Democracy and Technology: The E-government handbook for developing countries". .2002 s.l.:s.n.
[76] Rogers, E. M. "Diffusion of Innovations (4th Edition). New York: The Free Press." 1995."
[77] Davis, F. "A technology acceptance model for empirically testing new end-user information system: theory and results, Doctoral Dissertation, Sloan School of Management, Massachusetts Institute of Technology" 1986.
[78] Davis, F. "Perceived usefulness, perceived ease of use, and user acceptance of information technology, MIS Quarterly", 1989. 13(3), pp. 319-340.
[79] Davis, F. "User acceptance of information technology: system characteristics, user perceptions and behavioral impacts. International Journal of Machine Studies," 1993. 38(3), 475-488.
[80] Goodhue, D., and Thompson, R. "Task-technology fit and individual performance, MIS Quarterly, 1995 19 (2), 213-236.
[81] Alharbi, S., and Drew, S. "Using the Technology Acceptance Model in Understanding Academics' Behavioural Intention to Use Learning Management Systems, International Journal of Advanced Computer Science and Applications (IJACSA)" 2014, 5(1), pp. 143-155.
[82] Calisir, F., Gumussoy, C. A., and Bayram, A. "Predicting the behavioural intention to use enterprise resource planning systems - An exploratory extension of the technology acceptance model, Management Research News", 2009, 32(7), pp. 597–613.
[83] Sun, Y., Bhattacherjee, A., and Ma, Q. "Extending technology usage to work settings: The role of perceived work compatibility in ERP implementation, Information and Management", 2009. 46, pp. 351–356.
[84] ITU and OECD, "M-Government: Mobile Technologies for Responsive Governments and Connected Societies. ITU (International Telecommunication Union) and OECD (Organisation for Economic Co-operation and Development), Geneva, Switerland: International Telecommunication Union". 2011.
[85] ITU, "Least developed Countries, Commited to connectng the world." 2017.
[86] ITU, "ICTs, LDCs and the SDGs Achieving universal and affordable Internet in the least developed countries, International Telecommunication Union Telecommunication Development Bureau Place des Nations CH-1211 Geneva 20 Switzerland." 2018.
[87] ITU, "Committed to connecting the world; Least Developed Countries on track to achieve SDG 9.c on universal and affordable Internet access by 2020 Investment in digital skills and infrastructure still needed though to bring more people online." 2018,
[88] Richardson, J. W. "Technology Adoption in Cambodia: Measuring Factors Impacting Adoption Rates. Journal of International Development." 2009b. 21 (2): 1–14.
[89] Rogers, E. M. "Diffusion of Innovations. 5th Ed. New York: Free Press." 2003.
[90] Conde, M. Á., Garcia-Peñalvo, F. J., Rodriguez-Conde, M. J., Alier, M., Casany, M. J., & Piguillem, J. "An evolving learning management system for new educational environments using 2.0 tools. In Interactive learning environments." 2012.
[91] Cox, M. J. "Formal to informal learning with IT: Research challenges and issues for e-learning. Journal of Computer Assisted Learning", 2012. 29(1), 85–105.
[92] García-Peñalvo, F. J., de Figuerola, C. G., & Merlo, J. "Open knowledge: Challenges and facts. Online Information Review", 2010. 34(4), 520–539.
[93] Khaddage, F., & Knezek, G. "Convert your thinking! Creativity and imagination using mobile applications. In S. Dowling (Ed.), E-learning in action 1". 2012. Pearson/HCT Press, pp. 69–77.
[94] ITU, "ICTs, LDCs and the SDGs Achieving universal and affordable Internet in the least developed countries. International Telecommunication Union Telecommunication Development Bureau Place des Nations CH-1211 Geneva 20 Switzerland." 2018.

OTA-C Filters for Biomedical Signal Processing Applications using Hybrid CMOS-CNFET Technology

S.B. Rahane
MCOERC, Nashik
Savitribai Phule Pune University
Nashik, India
sandiprahane@ieee.org

A.K. Kureshi
VACOE, Ahemadnagar
Savitribai Phule Pune University
Ahemadnagar, India
akkureshi@rediffmail.com

G. K. Kharate
MCOERC, Nashik
Savitribai Phule Pune University
Nashik, India
gkkharate@yahoo.com

***Abstract*—Analog filters for biomedical signal processing applications deals with very slow or low frequency electrical activities of the physiological signals. This paper proposes first order, second order, fifth order elliptic low pass, second order notch and high pass OTA-C filters using hybrid CMOS-CNFET technology. Carbon Nanotube Field Effect Transistors (CNFETs) and CMOS devises can be heterogeneously integrated on a single 3-D chip to realize important signal processing building blocks such as OTA-C filters. Proposed circuits uses hybrid CMOS-CNFET Operational Transconductance Amplifier (OTA) as a building block for OTA-C filters. Realized filter circuits satisfy ultra-low power consumption requirement of wearable and implantable biomedical devices. The transistors used in the circuit operate in weak inversion to achieve ultra-low power consumption.**

Keywords—Carbon Nanotube Field Effect Transistors, OTA-C Filters, Noise Filtering, Biomedical Signal Processing.

I. INTRODUCTION

Filters preserve the signal of interest and at the same time systematically reject unwanted noise and interference. Real world signals including biomedical signals are almost always susceptible to the unwanted noise, interference and artefacts unless filtered properly. In case of biomedical signals, often the frequency spectra of signal of interest overlap with that of noise components and signal artefacts. This makes biomedical filtering a difficult task. Analog or digital filters can be used to achieve this task. Analog active filters are built around operational amplifiers (OPAMPS) or operational transconductance amplifiers (OTA). Amongst the different type of active filters OTA-C filters provide wide frequency range for the desired filter cut off frequencies. OTAs form the basic building block of OTA-C filters. OTA-C filters are better suited for on chip integration with a simple way to do frequency tuning either by changing OTA transconductance or by choosing suitable capacitor values [1]. The OTA -C filters find applications in many biomedical signal processing applications. Examples include Electrocardiogram (ECG) [2], bionic ear [3], neural recording interfaces [4], and electroencephalogram (EEG) [5]. The OTA-C filters are the better choice for analog front ends of biomedical systems involving low frequency physiological signals. In this paper we propose OTA-C filter designs based on hybrid CMOS-CNFET OTA. Following sections in this paper describe brief review of CMOS-CNFET co-integration (Section II), followed by hybrid CMOS-CNFET OTA (Section III), OTA-C filter realizations (Section IV) and simulation results with discussion (Section V).

II. A REVIEW OF CMOS-CNFET CO-INTEGRATION

CNFET is one of the promising alternative to the conventional CMOS technology in the sub 20nm regime while at the same time provide an opportunity for hybrid co-integration in a complementary manner. Owing to its superior material and electrical properties along with CMOS fabrication compatibility, CNFET/hybrid CMOS-CNFET circuits show enormous potential for the emerging nano-electronic circuits. Hybrid CMOS-CNFET monolithic 3-D integrated circuit has been successfully demonstrated by Stanford Researchers [6]. An ability to heterogeneously integrate CNFET logic, CNFET gas sensors, memory (RRAM) and CMOS logic was demonstrated on a 3-D chip using a low temperature fabrication process [6]. A 3-D architecture comprising CMOS circuits, 3-D stacked CNT-NEM (nano-electromechanical) switches and 3-D on chip cross bar bus is discussed in [7]. A CMOS nano-hybrid FPGA architecture consisting CMOS co-integrated with carbon nanotube crossbar memories and CNT bundle interconnects is presented in [8]. A cascode amplifier utilizing a hybrid combination of nMOS and pCNFET was used by researchers [9] as a vehicle to demonstrate hybrid CMOS-CNFET monolithic integration. Such type of hybrid architectures are promising approach for wearable and implantable biomedical devices.

III. HYBRID TRANSCONDUCTOR

Fig. 1 shows the transconductor (OTA) which is realized as a hybrid circuit with P type CMOS transistors and N type CNFETs for the OTA-C filters. Synopsis HSPICE circuit simulation tool is used to simulate the realized OTA-C filters. CMOS transistors are modelled as per Predictive Technology Model (PTM) 32nm low power device technology while the CNFET devices use Stanford University compact spice model [10, 11]. The transistors are sized as W=100, L=2 (PMOS) and (19, 0) chiral vector with 3 number of CNTs for N-CNFETs. CNFETs offer several benefits over CMOS with regards to higher drive current, lesser process variations and less leakage current. The output current expression of the transconductor is given by,

$$I_{out} = gm\,(V_p - V_m) \qquad (1)$$

Where gm is the transconductance of the OTA. The OTAs can only drive capacitive loads and do not require to bias the resistive elements and output driver circuits resulting in a circuit with both its inputs and outputs as very high impedance node. This makes the OTA to act as a voltage controlled current source. The OTAs of this kind are the circuits of

choice for biomedical signal processing applications such as neural amplifiers for electrocorticograms and electroneurograms[12].

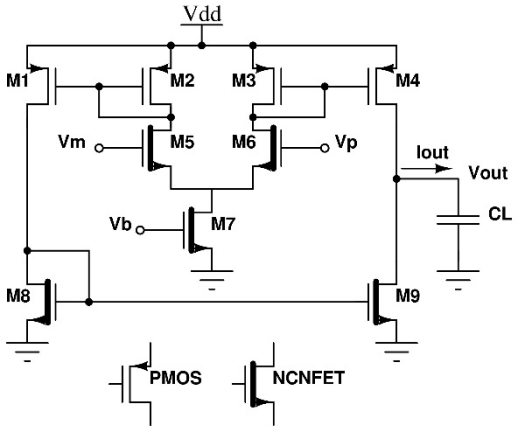


Fig. 1. Hybrid Transconductor

To minimize the OTA transconductance and improve the linearity of the transconductor transistors are forced to operate in weak inversion resulting in average power consumption of the proposed OTA to be just 4.45µW.

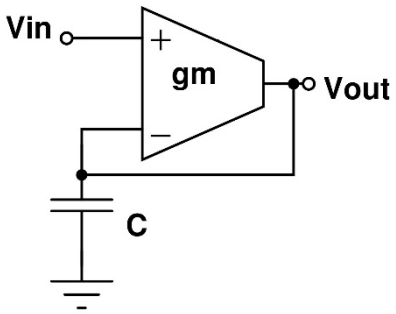
IV. HYBRID OTA-C FILTERS

Utilizing the hybrid transconductor shown in Fig. 1 different filter structures such as 1st order, 2nd order, 5th order elliptic low pass and 2nd order notch OTA-C filter structures are realized and simulated with HSPICE simulation tool. Filtering along with low noise amplification is an important signal conditioning requirement of typical biomedical signal processing chain. For most of the biomedical applications typical frequencies (cut off) of interest fall within few hertz to around 150Hz.For example typical patient monitoring filter cutoffs are- respiration-15Hz, EEG (Electro-encephalogram)-35Hz, EOG (Electro-oculogram)-35Hz, ECG (Electro-cardiogram)-70Hz, EMG (Electromyogram)-100Hz and snoring-100Hz. The frequency ranges for these physiological signals and their origins are given in [13]. The bandwidth characteristics differ for these physiological signals and typical filter cut off recommendations also vary depending upon the specific application. Diagnostic requirement of low pass filter cutoffs are different with recommendations as 15Hz for ECG,70Hz for EEG and 500Hz for EMG. Design of very low frequency filters for these applications has been a major concern due to the requirement of large valued passive components [14]. To achieve low cutoffs using OTA-C filters need large capacitors. Realizing large value capacitance/area using conventional CMOS technology poses several issues such as area overhead and parasitics. Even a gate channel capacitor approach which can offer considerably large capacitance/area has a limitation that the voltage across it should be larger than threshold voltage. This issue can be tackled with usage of VLSI-MOS compatible deep trench capacitors and Integrated Passive Device (IPD) technology. Both deep trenching and IPD approach allows integration of high capacitance/area in a monolithic chip [15-18].

First order realization of low pass filter using the hybrid transconductor is shown in Fig. 2. The desired cut off frequencies can be obtained either by adjucsting transcondcutance (gm) or by choosing appropriate value for the capacitor. The bias current of the transconductor decides the gm. The bias current is set to 0.1µA by appropriately sizing the CMOS transistors. A small value bias current results in very low frequency cut offs for intended application and at the same time reduces power dissipation significantly. Keeping the bias current as low as possible is one of the most efficient way to reduce power dissipation in OTA-C filters. The proposed filter structures used a fixed transconductance(gm) and are tuned electronically to achieve desired cut off frequencies by suitably choosing capacitor values.

V. RESULTS AND DISCUSSIONS

Frequency response showing the magnitude (gain) and phase plots for the first order OTA-C filter are shown in Fig. 3. The capacitance selection in the range of 1pF to 10000pF results in 3dB cut off frequency between 300 KHz to 30Hz. For example a 3000pF capacitor results in low pass 3dB cut off frequency of 100Hz. The average power consumption of the first order low pass filter circuit is 216nW.



$$\frac{V_{out}}{V_{in}} = \frac{gm}{sC + gm} \qquad (2)$$

Fig. 2. 1st order low pass OTA-C filter

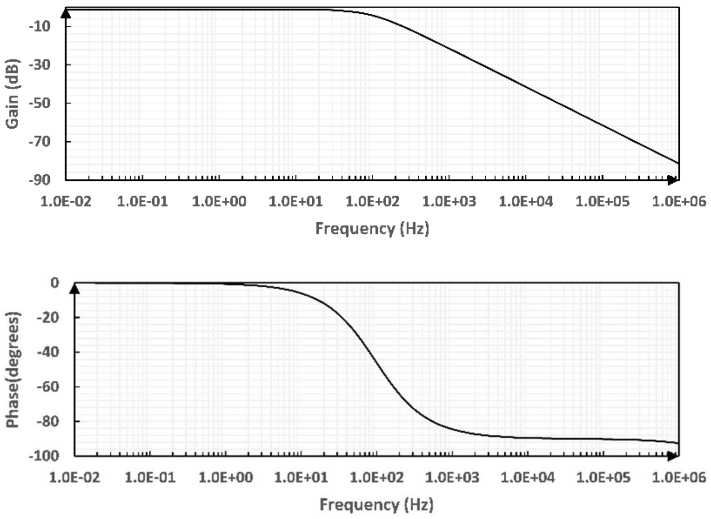
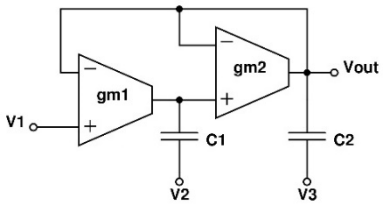


Fig. 3. Frequency response of 1st order low pass OTA-C filter



$$\frac{V_{out}}{V_{in}} = \frac{gm1gm2}{s^2C1C2 + sC1gm2 + gm1gm2} \qquad (3)$$

Fig.4. 2nd order low pass OTA-C filter

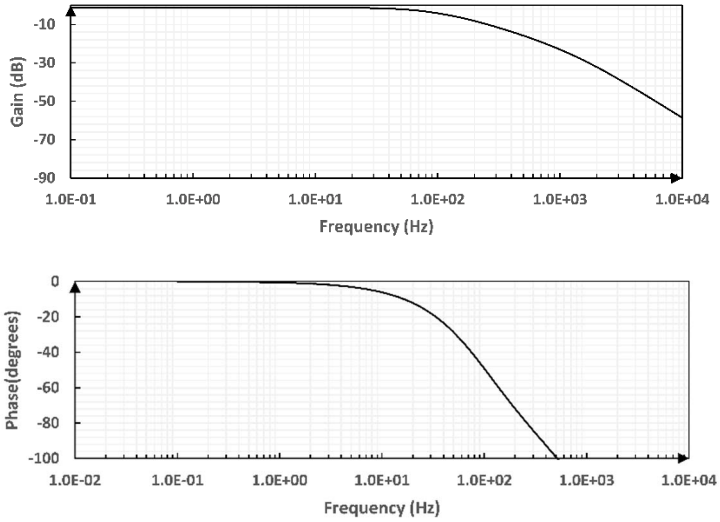


Fig.5. Frequency response of 2nd order low pass OTA-C filter

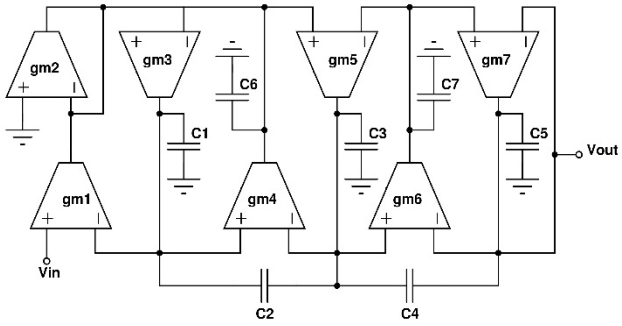


Fig. 6. 5th order elliptic low pass filter

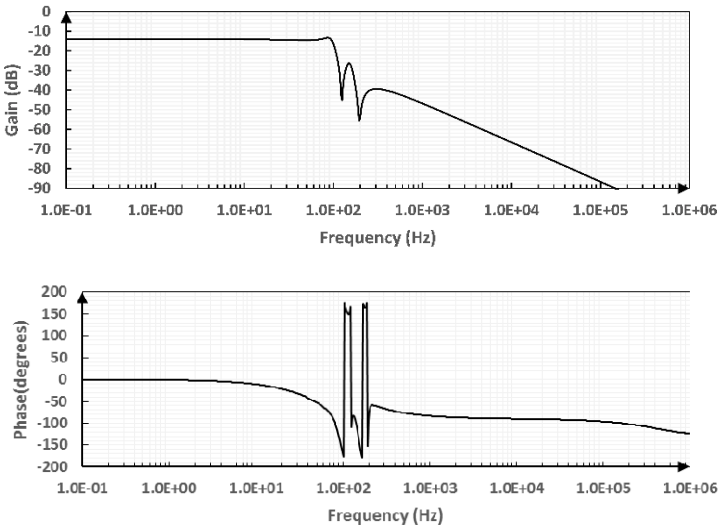


Fig. 7. Frequency response of 5th order elliptic low pass filter

Fig. 4 shows the configurable filter structure in its second order form (transfer function given by Eq. 3) and Fig 5. shows its frequency response. Different configurations such as low pass, high pass, band pass and band reject (notch) circuits are possible with appropriate inputs. A low pass realization is obtained when we provide input signal (Vin) at terminal V1 and terminals V2 and V3 are grounded. The output equation of the second order low pass OTA-C filter is given by,

$$V_{out} = \frac{s^2C1C2V_3 + sC1gm2V_2 + gm1gm2V_1}{s^2C1C2 + sC1gm2 + gm1gm2} \qquad (4)$$

For both the OTA's in the filter structure we keep the transconductances gm1 and gm2 equal while desired cut off frequency of 100Hz is obtained by choosing C1=3000pF and C2=200pF. Similarly to obtain lower cut off frequency of 30Hz, C1=500pF and C2=10000pF are used. C1=50pF and C2=4000pF are chosen to obtain 70Hz cut off. A 433nW power is consumed by this circuit. The second order filter is configured as 50Hz notch filter (Fig. 4) by applying input signal (Vin) to terminals V1 and V3 and grounding the terminal V2 with capacitance values of C1=10000pF and C2=3100pF. One cannot neglect 50Hz/60Hz power line frequency noise in case of bio potential measurements and hence a notch filter is an essential requirement in biomedical signal processing [19]. The frequency response showing a 50Hz rejection is given in Fig. 8 and the notch circuit consumes 431nW power. A low cut off (10Hz) high pass version can be realized with Fig. 4 by grounding terminals V1, V2 and applying input signal (Vin) to terminal V3. The frequency response of the high pass OTA-C filter is shown in Fig. 9 and the circuit consumes 45nW power. Such type of very low cut off is essential to eliminate DC and low frequency components introduced by skin-electrode interface in the physiological measurements. For example a surface EMG signal is often contaminated by the noise introduced at

the skin-electrode interface due to muscle movements and electro-chemical noise. Typical high pass filter cut off recommendations for this application range from 5Hz to 20Hz [20].

A more complex hybrid OTA-C filter realization is a fifth order elliptic structure exhibiting sharper roll-off is shown in Fig. 6. The faster transition from pass band to stop band or sharper roll-off is characterized by ripples in the pass band as well as in stop band. The elliptic filter dissipates 1.4714µW power. The frequency response (Fig. 8) with 100Hz cut off is obtained with capacitances C1-C5 equal to 2000pF, C6=2500pF and C7=1000pF.

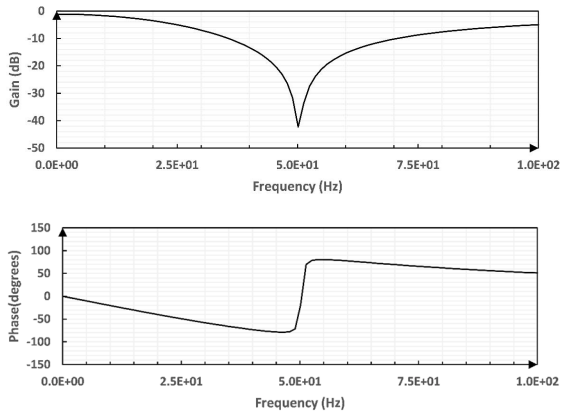


Fig. 8. Frequency response of OTA-C notch (50Hz) filter

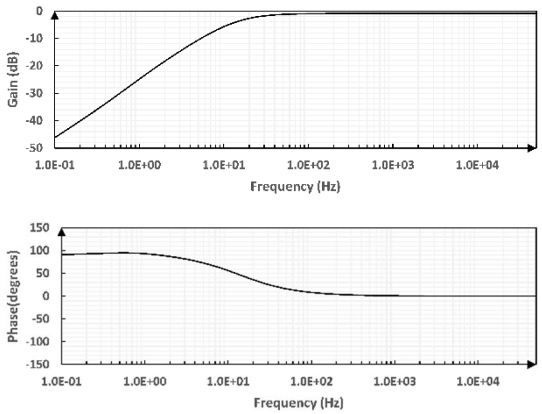


Fig. 9 Frequency response of OTA-C high pass filter

In this paper we have presented application of proposed filter structures for different filtering needs in complex and very low frequency physiological or biomedical signals. Noise removal with respect to 50Hz power line frequency noise (notch filter) and skin-electrode interface noise removal using a high pass filter is also discussed. However one needs to explore further and investigate filter designs that deals with different kinds of noise signals such as baseline wander, motion artifacts, noise due to muscle contractions and the internal noise generated by the signal processing circuit itself.

VI. Conclusion

OTA-C filters for biomedical signal processing, designed with hybrid CMOS-CNFET operational transconductance amplifier as a building block has been presented in this paper. CNFETS are proposed as an alternative to silicon MOSFET, but in near future it is very difficult to replace the CMOS technology because of research advancement and economic impact of CMOS based devices. To bridge the gap CNFET technology can be combined with proven CMOS technology resulting in a hybrid CMOS-CNFET circuits which can form the basis of variety of circuits and systems. Owing to difficulties in realizing the required large value capacitors in CMOS process one can opt for integrated deep trenching or integrated passive device technology for the proposed filters. In addition to capacitance and transconductance variation, desired cutoffs can also be achieved by choosing appropriate number of nanotubes.

References

[1] R.K. Ranjan, S.P. Yalla, S. Sorya, and S.K. Paul, “Active comb filter using operational transconductance amplifier,” Active and Passive Electronic Components, pp. 1-6, 2014.

[2] D. Jihai, L. Chuang, X. Weilin, and W. Baolin, “An OTA-C filter for ECG acquisition systems with highly linear range and less pass band attenuation,” J. of Semiconductors, vol. 36,no. 5, pp. 055006, May 2015.

[3] S. Peng et al., "A power-efficient reconfigurable OTA-C filter for low-frequency biomedical applications," IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 65, no. 2, pp. 543-555, Feb. 2018.

[4] J. Jang, J. Lee, N. Choi and Y.K. Song,"16 channel ultralow power analog front-end for an implantable neural recording microsystem," DEC J. Integrated Circuits Syst., vol. 02, no.1, pp.46-52, April 2016.

[5] A. Deo, S. K. Pandey, A. Joshi, S. K. Sharma and H. Shrimali, "Design of a third order butterworth Gm-C filter for EEG signal detection application," 2018 25th International Conference "Mixed Design of Integrated Circuits and System" (MIXDES), Gdynia, 2018, pp. 361-365.

[6] M. M. Shulaker, et al., "Three-dimensional integration of nanotechnologies for computing and data storage on a single chip," Nature, vol. 547, pp. 74-78, July 2017.

[7] S. Fujita, K. Nomura, K. Abe, and T.H. Lee, “3-d nanoarchitectures with carbon nanotube mechanical switches for future on-chip network beyond cmos architecture,” IEEE Trans. Circuits Syst. I: Reg. Papers, vol. 54, no.11, pp.2472-2479, Nov. 2007.

[8] C. Dong, D. Chen, S. Haruehanroengra, and W. Wang, “3D nFPGA:A reconfigurable architecture for 3D CMOS/nanomaterial hybrid digital circuits,” IEEE Trans. Circuits Syst. I, vol. 54, no. 11, pp. 2489-2501, Nov. 2007.

[9] D. Akinwande, S. Yasuda, B. Paul, S. Fujita, G. Close, and H. Wong, “Monolithic integration of CMOS VLSI and carbon nanotubes for hybrid nanotechnology applications,” IEEE Trans. Nanotechnol., vol. 7, no. 5, pp. 636-639, Sept. 2008.

[10] J. Deng and H. S. P. Wong, "A compact SPICE model for carbon-nanotube field-effect transistors including nonidealities and its application—Part I: model of the intrinsic channel region," IEEE Trans. Electron Devices, vol. 54, no. 12, pp. 3186-3194, Dec. 2007.

[11] J. Deng and H. S. P. Wong, "A compact SPICE model for carbon-nanotube field-effect transistors including nonidealities and its application—Part II: full device model and circuit performance benchmarking," IEEE Trans. Electron Devices, vol. 54,no. 12, pp. 3195-3205, Dec. 2007.

[12] K.A. Ng, E. Greenwald, Y.P. Xu, N.V. Thakor, "Implantable neurotechnologies: a review of integrated circuit neural amplifiers,” Medical and biological engineering and computing, vol. 54, no. 1, pp.45-62, Jan. 2016.

[13] J. G. Webster , Medical Instrumentation: Application and Design , 4th ed. Boston, MA: Houghton Mifflin, 2010.

[14] C. Sawigun and W. A. Serdijn, "A modular transconductance reduction technique for very low-frequency Gm-C filters," 2012 IEEE International Symposium on Circuits and Systems, Seoul, 2012, pp. 1183-1186.

[15] N. Habib, M. Muhammad, J. P. Bickford, J. Safran, A. Y. Ginawi and F. J. Towler, "Deep trench capacitor in three dimensional through silicon via keepout area for electrostatic discharge protection," IEEE Trans. Semicond. Manuf., vol. 29, no. 4, pp. 292-298, Nov. 2016.

[16] A. Aminulloh, V. Kumar, S. Yang and G. Sheu, "Novel structure of deep trench capacitor with higher breakdown and higher capacitance density for Low Dropout Voltage regulator," 2013 IEEE 10th International Conference on Power Electronics and Drive Systems (PEDS), Kitakyushu, 2013, pp. 389-392.

[17] B. Kim, W. Xu and C. H. Kim, "A supply-noise sensitivity tracking PLL in 32 nm SOI featuring a deep trench capacitor based loop filter," IEEE J. Solid-State Circuits, vol. 49, no. 4, pp. 1017-1026, April 2014.

[18] L. Sun, Y. Chen and K. Sun, "System integration using silicon-based integrated passive device technology," 2012 IEEE International Symposium on Radio-Frequency Integration Technology (RFIT), Singapore, 2012, pp. 98-100.

[19] M. S. Ansari, R. Sinha and S. Khot, "Ultra-low power 50/60Hz notch filter for biomedical signal acquisition using 32nm ± 0.15V Bulk-Driven subthreshold CMOS OTAs," 2017 4th International Conference on Electrical and Electronic Engineering (ICEEE), Ankara, 2017, pp. 309-313.

[20] R.H. Chowdhury et al., "Surface electromyography signal processing and classification techniques," Sensors, vol. 13, no.9, pp. 12431-12466, Sep 2013.

Resource Leveraging and Teaching Computing in the New Lower Secondary Mauritian Curriculum

Nirmal Kumar Betchoo
Dean of Faculty
Department of Business and Management
Université des Mascareignes
Beau Plan, Pamplemousses
nbetchoo@udm.ac.mu

***Abstract*—The Ministry of Information Technology, Information and Communication, under the aegis of the Mauritian government aims at developing digital learning to improve the potential of Mauritian students for the future. As the country posits itself as a cyber-island with a fully developed information technology hub, it has become imperative to think about leveraging the teaching of computer science in lower secondary schools (Grades 7-9). So far, this has been limited to basic courses in computing that are essentially theory focused. This paper states that students at this stated level need to develop 'computational thinking' so that they are ready for the workplace and will effectively enter the digital world. It limits itself to two key variables resource leveraging and teaching computing. In the first case, it raises arguments on programming languages, hardware and software that might apply to students in the new curriculum. In the teaching area, the paper suggests that a combination of teaching programming, inclusion and informal learning will be the possibilities to accompany students overcome the digital divide and get the opportunity to become effective as learners and trained ICT students to serve their nation.**

Keywords—resource leveraging, teaching, programming, software, digital divide, lower secondary, Mauritius

I. Introduction

There is an intention of making Mauritius, a republic island-nation in the Indian Ocean, make some quantum leap in its Information and Communication development strategy in the years to come. So far, the nation has managed it positively but there needs to be deep thinking regarding developing computer sciences at younger age for the immediate benefit of students so that they are better prepared for tomorrow's challenges. Very often, it is at the end of the secondary education or when joining universities that higher-end learning in computer sciences is developed. This paper considers the problem at the early years of secondary education which could prompt potential students learn computer sciences both with literacy and practical elements. The combined prowess could enable the young generation to become more apt to develop competences in computer sciences and enable them become more efficient learners and practitioners in the future.

The Government of the United Kingdom in its National Curriculum in England stated that "a high-quality computing education equips pupils to use computational thinking and creativity to understand and change the world [1]." It is a known fact that computing is linked with Mathematics, Science and Technology as well as design. From experience, it can also be assessed that computer science can be considered as the core of computing? This is the subject area where students learn about principles of information and computation, how digital systems work and how to convert such knowledge into programming.

Building on the foundation of computer knowledge and understanding, students in secondary schools could be equipped with such talent to use information technology to create programs, systems and even a range of content and practical applications in their studies. In the Mauritian context, the national ICT policy states that information technology will be taught in schools as a subject and integrated into teaching across the curriculum in primary and secondary schools. The challenge is to promote connectivity in schools and establish a network for information exchange and information in the education sector [2].

The coming sections will develop the problem statement regarding the need for implementing computer education in lower secondary schools, followed by a literature review, value propositions, the case for resource leveraging and the relevance of teaching of computer sciences in the appropriate way in the learning environment.

II. Problem Statement

The case in Mauritius deserves interest as there is an effort undertaken by the educational authorities to impart computer education at all levels in the stages of child development until adulthood. This is commendable in the sense that computer education is fostered since childhood but, it is seen that, in the long-run, it benefits a handful of capable IT students at the university level to master programming and to use computer sciences to develop applications and software that are needed by their employers or in industry.

External curricula in advanced countries claim that there must be some effort done in lower secondary classes to ensure that there is a 'kick-start' to learning computer science and that this might be a trigger to more effective computer learning compared to theory that is initiated at the beginning.

To this end, the United Kingdom might be the first country in the world to mandate computer programming in schools. At three stages of learning, the following developments might have occurred. At stage 1, children might have the ability to create and debug programs as well as 'use technology safely and respectfully.' At stage 2, they learn to use the websites, store information and even share it. She adds that in stage 3 (around 11-14 years) that students will be taught how to design and write programs that accomplish specific goals, including controlling or stimulating physical systems [3].

They will also understand computer networks and use logical reasoning to detect and correct errors in algorithms. This is not just an evolutionary change-this is a massive revolution in the study of computing, which so far, consisted almost entirely of lessons on how to use Microsoft office programs.

This is where Mauritius direly requires a change in mind-set in its curriculum design to enhance its students' learning and adaptation to computer sciences at a younger age than waiting for a longer time.

Below is provided a comparative table with the KSA's formulated in England and how they could be mapped onto the Mauritian curriculum.

TABLE I. COMPARATIVE KEY STAGES IN THE UK AND MAURITIUS UNDER KEY STAGES OF LEARNING COMPUTER SKILLS

Comparative Key Stages		
Key stages	***UK***	***Mauritius***
Key Stage 1: Children learn to create, store, manipulate and retrieve digital content.	5-6 years	8-9 years
Key Stage 2: Children can learn to navigate through websites and use internet services.	7-11 years	7-11 years Grades (1-6)
Key Stage 3: Children can use one or two programming languages and create their own programs.	11-14 years	12-14 years Grades (7-9)

III. Literature Review

To be able to have some grasp of literature on computer training for students, some scholar contributions are hereby analysed.

Three basic questions could be asked regarding the pedagogical content knowledge that could be applied to young learners: why teach a subject, who should be taught and what are the learning difficulties? A basic argument here could be why a student should learn programming thought be reserved for a happy few intellectually-apt students [4].

In learning to program one learns about problem solving, design and thinking abilities. When they program, they need to find a solution to a problem, reflect on how to use and communicate with the machine, using programing syntax and grammar [5].

In addition to that, problem-solving skills can be deployed to solve 'realistic' problems together with computer goals. Mulder's contribution is important here where he states that programming is a new generation subject which brings together pieces from different areas such as linguistics, mathematics and economics [6]. Such completeness offers the students the opportunity to gain multi-disciplinary skills [7].

The challenge then comes from teaching advanced computer skills in the secondary classes and this is where the problem arises. The essence of programming is problem solving and developing a program as a solution [8]. There might be two kinds of knowledge namely program generation and comprehension. In program generation, the student analyses the problem, produces an algorithmic solution and then translates the algorithm in a program code [9]. Regarding program comprehension, the young programmer is expected to describe how the program works.

The problem comes from teaching difficulties in this quite challenging context. Programming is a difficult task to achieve and that novice programmers do show misunderstandings and misconceptions [10]. There is apparently no short-fix formula to learn. This is where adaptations need to be possibly taken from international environments to see how well some advancement in computer sciences could be useful if they were introduced in lower secondary schools.

IV. Value Propositions

Seen from the existing perspective, it might be useful to review some value propositions regarding the implementation and adoption of computer sciences in the Mauritian pedagogical system as a means of resource leveraging in the expected future. Is it possible to re-engineer the syllabus and the content to make it more adaptable and applicable to the Mauritian situation?

A. A Curriculum that is valuable and applicable

The UK National Curriculum for Computing (2018) aims for the following: (a) Students should be able to understand and apply the fundamental principles and concepts of computer science, including abstraction, logic and data representation. (b) Students can analyse problems in computational terms, and have repeated practical experience of writing computer programs in order to solve such problems. (c) Students possess the ability to evaluate and apply information technology, including new and unfamiliar technologies, and analytically solve problems and (d) such students are responsible, confident and creative users of information and communication technology [11].

B. Teaching Coding in Schools

There is merit in school students learning coding. In a digital world, computer programs underlie everything from business, science to several other ones [12]. Australia's prosperity will depend on delivering advanced services and digital technology, and that programming will be essential to this end. Computer programs and software are known as strong drivers of productivity improvements in many fields.

In Finland, coding is considered as a mandatory, cross cultural theme starting from first grade or primary. It positions coding as a new element to complement reading, writing, drawing and calculating. The Finnish approach could be worthwhile to other nations in seeing how coding enhances ICT learning [13].

C. Implementing a Teaching Policy on Programming

Although intentions are good, they need to be implemented in order to expect progress from them. For instance, the problem is tantamount in Nigeria where there is a wide disparity between policy pronouncements and implementations [14]. To address such problem, there were arguments the emergence of a 'policy elite' in the form of a lobby group to the government to enforce policy implementation [15].

V. Resource Leveraging In Computer Science

To address the current problem in Mauritius, it is essential to firstly speak of resource leveraging. To this date, the Government of Mauritius proudly advocates the country as a cyber-island with the enormous potential of becoming a hub for the sub-Saharan African region and the Indian Ocean. Mauritius might have taken the lead since the 1980 with the introduction of personal computers (Commodore 64, Amstrad), gone through the use and vulgarisation of personal computers in the 1990s, the use of laptops and notebooks with the advent of the Internet in the new millennium.

In the turn of this 2nd decade of the millennium, the wide availability and use of mobile applications, data and resource and the development of new concepts like 'The Internet of Things' might have called for a leveraging of resources and competences in the ICT sector.

This has also raised the question of computer education and where should it start? Apparently, it is not a simplistic perception that children become 'digital natives' just by using the mobile but must develop an adaptive learning as they develop so that they are capable of harnessing their skills through the use of computers.

The framework proposed in this paper then analyses two key variables that will be developed: Firstly, resource leveraging through the development of programming languages, hardware and software applicable to lower secondary school students and, secondly, a mix of teaching programming, inclusion and informal learning that might help in overcoming the digital divide that affects existing learners.

VI. Resourcing: Three Suggested Options

There are possibilities of leveraging resources in the teaching of computer sciences. Programming languages, hardware and software should be included at this level. These are the basics needed for effective learning. They are outlined below.

A. Programming Languages

Taking into consideration that the students of lower secondary level are being considered in this study, there might be strong arguments to use the major programming languages at Key Stages 3 (KS3). Such teaching of programming should help students get the underlying skills of sequence, selection, repetition.

There is a broad range of languages available and the user/instructor can pick at least one that is textual. Visual language such as Scratch and Kodu are suitable starting points but there must be assurance that students have prior knowledge of them [16]. A refresher course might help here. There might also be general-purpose textual languages like Python, Visual Basic.NET and Java that allow for a wide range of project work, graphics and applications. It might be also recommended the language of spreadsheets as a textual language providing insight into functional programming. Javascript is a fully-featured programming language and can be effectively used in teaching programming.

B. Hardware

Colleges in Mauritius could provide the hardware to students as there are dedicated computer labs in each secondary school. Licensed programmes as those mentioned above could be well installed in the Personal Computers. There is a possibility to consider Raspberri Pi as an excellent teaching tool since it offers access to operating system functionality [17]. The command line functionality allows students to give the computer textual commands and query details such as network settings. It is expected that students should also learn to be conversant with appliances like small video camcorders, online servers, sound recorders, mobile phones and tablets.

C. Software

When it comes to using software in lower secondary schools, a range of such facilities might be available. In an exhaustive way, a few possibilities are hereby suggested for lower secondary school students.

Codeacademy might be considered as an educational resource that offers a variety of programs and applications that helps teachers easily integrated a variety of programs and applications in the computer science curriculum [18].

Scratch is another useful programming language designed for the target group 8-16. Students might learn how to program interactive games, stories and animations while sharing their projects. Scratch is believed to be an effective way to teach secondary school students creative thinking as how well to collaborate and use reasoning to complete programming projects [19].

Codio is a complete platform solution designed to support secondary computer science teachers by offering a cost-effective alternative solution for teaching computer science. Codio provides the instructor monitoring tools which allow him/her to track individual progress and manage classes well. Immediate feedback on student progress can be envisaged with this software [20].

Other software suggested might be: Codebug, Code HS, Python, Appshed.Acacemy, CodeOrg. Studio. These could be consulted from the websites.

VII. Teaching of computer programming

The second key consideration of this research paper is the teaching of computer programming in lower secondary schools. Three concepts are briefly outlined herein: Teaching practice, informal learning and inclusion.

A. Teaching Practice

Based on Papert's theory of learning known as constructionism, people learn best through making things for other people [21]. Students will then learn to become better when they are aware that others will read their work. They aim under constructivism to adopt computer science better when they write programs and create theory questions to teach others.

Project work, sharing of activities, blogposts, screencasts are useful ways of enhancing teaching. The sharing allows other users, students and instructors to question the quality of work and make improvements as and when required.

Since teaching practice is the main issue here, instructors are expected to understand the abilities of the different

students, coach them effectively, prompt low learners but also ensure that progress is diligently followed so that they are no laggards or failures within the teaching exercise. There is here the possibility of introducing role play, creative writing and the use of pencil and paper to clarify misunderstanding and permeate good learning approach.

B. Informal teaching

Informal teaching is also a good component to foster learning whether it pertains to secondary students or vocational learners. Much software in computer programming is available free from the web; some software companies offer computer licences specially for students. Examples might be YouTube, MIT ScratchEd for educators using Scratch with their students.

Teachers must know how to structure and enable learning. Allowing students learn through exploration, create, change and make personal interventions in their learning could be very useful ways of improving informal learning and enhance learning with computers.

C. Inclusion

In this digital age, the existence of the web, interactive whiteboards, virtual learning environments, video conferencing, blogs, podcasts and mobile devices have made a transformative impact on both teaching and learning. The digital divide must be understood from the point of view that all students might not have access to formal education in computer sciences in Mauritius but there might be possibilities to promote such social learning vocationally.

Inclusive education is about reforming schooling as a means of supporting education for everybody while removing barriers to participation and learning for disadvantaged groups. These link education reform with policies to alleviate poverty, improve child health, promote gender equality, environmental sustainability and global partnership [22].

Running of classes after schools, having computers and software in libraries and, more importantly, using licensing agreements or open source software might be promoters of social learning. These forms of assistive technologies could also address students with special educational needs or disabilities. In this concern, there must be an effective way of teaching and this is achieved by formulating computer knowledge in a way that it can be understood by students [23]. This is a possibility to improve such competences provided that learning computer programming starts from program comprehension and ends with program generation [24].

VIII. Conclusion

This research paper posited the importance of leveraging resources for lower secondary school students in Mauritius in the age group 12-14 within the Key Stage 3. Taking into consideration the need for the Mauritian government to raise the level of competence of such a target group and enable them learn deeper Computer Sciences compared to a traditional approach today, it is clearly felt that a creative methodology must be developed. To create a suitable model for learning, two variables were discussed namely resource leveraging through tailor-made programming languages, adequate hardware and dedicated software programs for the targeted audience. Secondly, the teaching of computer programming using Papert's model, informal learning and inclusion might be effective enhancers of computer science learning in the lower secondary schools of Mauritius.

Acknowledgment

Papert's Constructivism model has been used as a framework for this research work based on its international application and success obtained in the United Kingdom and Australia. The UK Department of Education Key Stage model has been referenced in Table 1.

References

[1] Government of United Kingdom (2018) National Curriculum in England, Computing Programs of study.

[2] S.Isaacs (2007) ICT education in Mauritius, Survey of ICT in Mauritius in Africa, Mauritius Country Report.

[3] S.Curtis. (2013) Teaching our children to code: A quiet revolution, The Daily Telegraph, 4th November 2013.

[4] P.Grossman (1989) A study of contrast: Sources of pedagogical knwledge, J.ournal of Teacher Education, 24-31.

[5] P. Szlávi, P and L. Zsakó. (2006). Programming versus application. In: Mittermeir, R.T. (Ed.), *ISSEP 2006, LNCS.*

[6] V. Dagiene. (2005). Teaching information technology in general education: challenges and perspectives. *Proceedings of International Conference on Informatics in Secondary Schools – Evolution and Perspectives*, Klagenfurt, Austria, 53–64.

[7] F. Mulder. (2002). English: Computer Science: from a BÈTA to a DELTA subject). *Informatica, Tinfon*, 11, 48.

[8] R. Romeike. (2008). What's my challenge? The forgotten part of problem solving in computer science education.

[9] L. Mannila, L. (2007). Novices' progress in introductory programming courses. *Informatics in Education*, 6, 139–152.

[10] I. Govender. (2006). *Learning to Program, learning to Teach Programming: Pre- and In-service Teachers' Experiences of an Object-oriented Language*. University of South Africa.

[11] Government of United Kingdom (2018) National Curriculum in England, Computing Programs of study.

[12] L.Sterling (2015) An education for 21st century means teaching coding in schools, Swinburne University of Technology.

[13] T.Toikannen (2015) Coding in School: Finland takes the lead in Europe, legroup.aalto.fi.

[14] H. Jeter. (2002): Computer, Technology and Office equipment show (CTO 2002): Keynote address.

[15] O. Maduekwe. (2003). Retrieved August 31, 2018., from the World Wide Web: *Maduekwe Seeks Emergence of "Policy Leadership" This Day* http://fr.allafrica.com/stories/200303030320.html.

[16] Stemlearning (2015) Coding using Scratch and Kodu, www.stem.org.uk.

[17] RaspberryPi.org (2018) Coding Projects: Code clu bs, raspeberrypi.org.

[18] Ourict. (2015) Why choose Ourict? www.ourict.uk.

[19] MIT MediaLab (2018) Scratch: Create stories, games and animations, https://scratch.mit.edu.

[20] Codio.com (2018) Higher performance computer science programs, https://codio.com.

[21] S. Papert (1993). The children's machine: Rethinking schools in the age of the computer. New York: Basic Books.

[22] S. D'Alessio, V. Donnelly, and A. Watkins, (2010) Inclusive education across Europe: the move in thinking from integration for inclusion. Revista de Psicología y Educación, 1(5) p109-126.

[23] J. Loughran., P. Milroy., A. Berry., R. Gunstone, P. Mulhall, (2001). Documenting science teachers' pedagogical content knowledge through PaP-eRs. *Research in Science Education*, 31, 289–307.

[24] M.C. Linn, and J. Dalbey, (1989). Cognitive consequences of programming instruction. London, Lawrence Erlbaum Associates, 58–62.

An Access Control Framework for Protecting Personal Electronic Health Records

Ambrose Atuheire Izaara
Department of Information Technology, Faculty of Computing and Informatics Mbarara University of Science and Technology
Mbarara, Uganda
aturose@must.ac.ug

Richard Ssembatya
Department of Computer Science, Faculty of Science and Technology Uganda Christian University
Kampala, Uganda
richssembatya@gmail.com

Fred Kaggwa
Department of Computer Science, Faculty of Computing and Informatics Mbarara University of Science and Technology
Mbarara, Uganda
kaggwa_fred@must.ac.ug

***Abstract*—The increasing expansion of wireless systems and the extensive popularity and usage of mobile devices such as mobile phones and wireless tablets represents a great opportunity to use mobile devices as widespread health data access tools. Unfortunately, some problems impeding the general acceptance of mhealth such as privacy protection, limitation of wireless networks and handheld devices are still common. Challenges such as unreliable data repositories and limited connection speeds in resource limited environments are also evident. The inadequate capabilities of hand-held devices and wireless systems make these Public Key Cryptography based frameworks unsuitable for mobile networks. Moreover, these protocols were designed to preserve customary flow of health data, which is vulnerable to attack and increase the user's risk. This research drew its foundations from literature and theoretical review and used qualitative approaches. In this paper, the researchers build on existing concepts of Medical Information Systems and use of Symmetric Key Infrastructure to design a framework for secure access to personal electronic health records. The framework provides identity protection for a patient from all forms of unathorised data access. The framework not only reduces the computational operations between the engaging parties, but also achieves privacy protection for the user. Validation results from ICT experts demonstrate that the designed framework is applicable for secure access to personal medical health records in resource limited settings.**

Keywords—mhealth, security, framework, electronic health records

I. INTRODUCTION

The healthcare industry has received an explosion in innovation in the last decade aimed at improving life expectancy, quality of life, diagnostic and treatment options, as well as the efficiency and cost effectiveness of the healthcare system [1]. Opportunities of using mobile technology have improved in the last few years with the growing population of smart phones in Uganda [2]. A combination of wireless mobile technology and medical health innovation can revolutionarise the use of medical technologies to improve service delivery in the health sector. Personal Health Record (PHR) systems play a significant role in improving the efficiency of healthcare service delivery by enabling patients to monitor, control and provide vital information to healthcare givers. [3], [4]. Despite the many benefits of having PHR systems, the development of suitable and scalable PHR systems in emerging economies is still difficult due to limitations that are inherent to the technological and social issues [3]. As such, the mainstream of existing PHR practices in emerging economies are primarily paper-based [3]. Paper-based PHR systems are prone to loss and unavailability of patient information, delays in retrieving the information and space limitations for record-keeping [3]. Similarly, due to limited resources, such as intermittent power and Internet connectivity, computerizing these processes securely is challenging.

Furthermore, literature shows a number of studies demonstrating that PHR systems implemented in emerging economies do not sufficiently protect patients' records [3], [5]. These findings were also in agreement with our study conducted at Mbarara University Teaching Hospital (Uganda). Most of the hospital personnel including doctors, nurses, lab technicians and messengers have access to all the personal health records for all the patients. This demonstrates that there is no role-based access control hierarchy implemented at the hospital, and therefore no mechanism of controlling authorizations of data access. Patients expressed fear of their health data confidentiality and showed a need to access their health data using personal mobile devices.

II. RESEARCH MOTIVATION

There is an increase in the use of hand-held devices in developing countries [3], [6], and user demands for more patient controlled access to healthcare data [5], [6,] [16]. Additionally, the growth of wireless infrastructure in developing countries has increased demand for mobile PHR systems [6], [16]. While mobile phones and the increasing availability of wireless infrastructure can support and improve patients' access to their personal Health Records, they generally do not provide sufficient security mechanisms for adequately protecting personal data from unauthorized disclosure, especially when patient's data is stored on the mobile device [3]. This is mostly due to the architectural deficiencies of their design [18]. The limited processing and memory capabilities of mobile phones to support resource intensive security architectures is still a major challenge. When personal data is transacted, downloaded and stored on mobile phones, it remains prone to being accessed by unauthorized parties. Therefore, an access control framework that securely protects mobile phone-based PHR is needed to facilitate secure sharing of Personal Health records.

III. RELATED WORK

A. Personal Health Records (PHRs)

In the current century, trends have favored PHRs to the customary Electronic Health Records (EHRs). PHRs allow users to be fully in charge of their own health records [3]. Unlike EHRs where providers have higher control, PHRs endow users to become the overseers, and have complete control of their health records.

Literature reveals two sorts of PHRs. These are paper-based PHR and electronic PHR. Paper-based PHRs are generally hard to share between healthcare givers and in many cases costly to transport [21]. Moreover, according to the medical record ethics, patient records should be kept for a definite number of years and should be readily available in order to improve patient care [11]. Keeping paper-based records for a number of years incurs high storage costs in addition to difficulty in interpretation of standard medical terms and jargon. This could result in deficiencies in terms of accuracy, availability and legibility of the stored documents [12].

A good way to overcome such deficiencies is to adopt the use of Electronic Personal Health Records (EPHRs). An EPHR is created by gathering health information of a person from a single or multiple source such that information can be shared through electronic networks with authorized healthcare professionals [3]. EPHR systems have been developed that offer patients with safe access to manage their own health information. However, recent studies reveal that there is no standard framework for EPHR [3], [13]. In the patient-centric EPHR model, patients manage their entire EPHR through web portals or hand-held computing devices such as mobile phones in order to access, read and update their records.

Mobile phone-based EPHR systems, due to their offline nature provide patients with a way to communicate with their healthcare givers when the hospital systems are offline, in addition to the provision of up-to date health records [3], [6], [16]. This makes a mobile phone-based EPHR more adequate in healthcare service delivery [3].

B. National Institute of Standards and Technology (NIST) Framework for Role Based Access Control

RBAC has four elements: users, roles, permissions and sessions; NIST RBAC elaborates permissions by introducing operations and objects sub entities [14]. In flat RBAC, users are assigned to roles, permissions are assigned to roles so users inherit permissions from being members of these roles. RBAC has default rights for users based on their roles. This implies that a doctor for example inherits rights that the doctors' role is defined. This access approach by itself does not satisfy requirements for Health Insurance Portability and Accountability Act (HIPAA). Therefore, basing on the framework that shall be designed, this approach will not be favourable as every patient needs to have only one role towards his data and a doctor who inherits another doctor's role may compromise the confidentiality of a patient.

C. Multi-Level Security Framework

MLS defines that every data has a classification and every user possesses a clearance. The security levels are unclassified, confidential, secret and top secret which are hierarchical. Multi-level security leverages on Bell La Padula security model [6]. Based on the Bell La Padula model, MLS permits users or processes to read only information classified at only or below their clearance. The fact that MLS is based on classification of information and clearance of the users prior to authorization limits its use in situations where the information sharing parties do not have prior knowledge of each other. In cases where a user is visiting a new medical practitioner, this would present privacy challenges.

D. Public Key Infrastructure (PKI) based Access control framework

PKI is a security architecture that uses the concept of a trusted third party to ensure confidentiality, integrity, non-repudiation and accountability during information sharing [16]. PKI is based on asymmetric cryptography where a pair of keys; public and private keys are used. What one key encrypts, only the other key can decrypt. The public key is published to the public while the private key is kept secret by a user. A sender uses the receiver's public key to encrypt data that can only be decrypted by using the receiver's private key. This provides for data confidentiality. PKI operations are very resource intensive and pose a challenge in resource limited environments. Therefore, with the environment for the proposed architecture to function well, PKI would not suit best.

E. Attribute Based Access Control (ABAC)

Similar to the above approach, the ABAC uses attribute certificates only that these certificates do not contain a public key [17]. An attribute certificate contains the account holder's specific attributes similar to policies that specify his or her access control information such as role, security clearance or group membership [17]. ABAC is effective at authorization of users from different security domains using a trusted third party even in environments where the parties do not have any prior knowledge of each other. ABAC does not use public keys and therefore does not cater for security during transmission. For the kind of data transmitted between health center and patients, the structural arrangement may not protect the user's privacy if data is intercepted on a network.

F. Surrogate Trust Negotiation (STN)

STN allows two parties that are previously unknown to each other outside a local security domain to transact securely through a handshake like process of requesting and providing digital credentials and policies [18]. "Trust agents are autonomous software modules on secure, offsite computers that act as surrogates for mobile devices, performing cryptographic operations and managing credentials, policies, secret keys for use in trust negotiation." [18] STN allows resource limited devices to participate in trust negotiation using trust agents.

G. Medical Information System (MEDIS)

MEDIS was developed in Serbia aiming at building an integrated patient centric electronic health record right from the beginning and not integrating existing heterogeneous ones [19]. It is designed to meet international health system standards such as Health Level 7 (HL7).

MEDIS has been implemented as a federated system where the central server hosts basic electronic medical records about the patient and the distributed clinical servers contain their own part of the patient's record [19]. MEDIS combines the strengths of the previous approaches to bring forward a federated system where the records are stored where they are collected but are accessible globally. It ensures security on the network using certificates, security from the client device using applets and conforms to public standards such as HIPAA. The research builds on this existing MEDIS literature for secure sharing and accessing E-PHR.

IV. PROPOSED FRAMEWORK

Although PHR systems promote the sharing of personal information between a patient and his/her healthcare provider, they also generate security and privacy issues [3] [22]. The consumer survey of PHR systems conducted by Markle Foundation working group found that ninety-one percent of the respondents are worried about the confidentiality of their PHR's. In the following section, an adversary model is presented and then a description on how the Framework operates is given.

A. Adversary Model

The framework design puts into consideration an adversary setting that defines the capabilities of possible actions of an attacker. Consideration of a probabilistic man in the middle attacker that has access to the communication links or the communication devices is made.

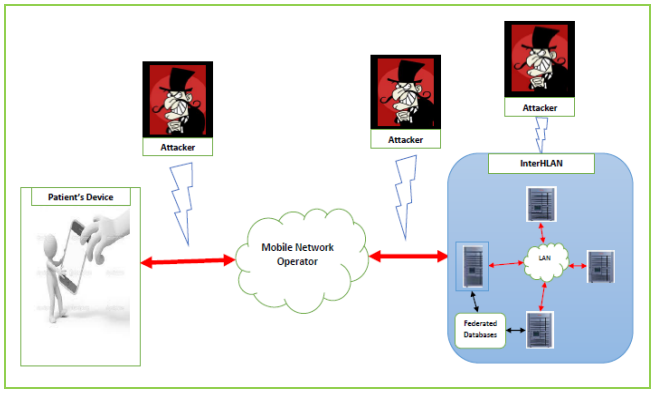


Fig. 1. Adversary Model

The attacker can listen to all transmitted messages or view transmitted messages on the communication device, change the EHR or inject their own generated messages. The Framework considers that all secrete information stored on a federated database or received on a mobile device is potentially vulnerable to break-ins or other forms of leakage. However, the framework guarantees that information leaked in one specific data transaction will have minimal effect on the security of other data transactions. The patient's identity is protected by use of random generated numbers sent to the federated data server during data transaction in case of a replay attack. Use of shared session keys between the Patients device and the federated data server verses a shared session key between the Mobile network operator and InterHealth Institution Local Area (InterHLAN) Network protects patient identity from a Trusted Third-Party attack. These keys are first exchanged during the registration protocol which only takes place on the InterHLAN to protect against eavesdroppers on public Mobile Networks.

To protect the privacy of the patient and resolve the problem of data flow between the user and the data repository, the framework is designed centered on a patient centric model. The Patients Data Server (P_{DS}) does not communicate with the Patient's Mobile Network Operator (MNO) directly. The framework is composed of four parties including Patients Device, Patients Data Server, Patient's MNO, and InterHLAN. The Framework works with the assertion that secrete Xi, where i=1...,n is only shared between Patients Device and Patient's MNO, and secrete Yi is only known between the Patients Data Server and InterHLAN. The following symbols are used in the framework.

TABLE I. NOTATIONS

SYMBOL	DESCRIPTION
{P_D P_{NO}, P_{DS}, InterHLAN}	A set of engaging parties which are The Patient's Device, Patient Device Network Operator, Patients Data Server and Inter Health Institutions LAN respectively.
TSC	Time Stamp Center
PN_p	SIM Number of Patients Device
PIN_p	Password Identification Number for Patient
ID_p	Identity of Patient which Identifies Patient to NO; computed as $ID_p=PN_p + H(PN_p, PIN_p)$
Bio_P	A set of the Patient's finger print scans to uniquely identify the Patient in emergency cases
AI_p	Account Information for Patient such as account status, age, names, address etc.
Rand	Random Number + Timestamp generated by P_D to avoid replay attack from a completed communication.
R	Random Number and Timestamp generated by P_{DS} to act as a P_{DS} ID which uniquely identifies P_{DS} to InterHLAN.
DATE	Date of Data Retrieval
ehr	Data to be retrieved
DESC	Description of data retrieved which may include delivery address. Patient will only disclose information they wish to disclose for privacy purposes.
TID	Identification of Data Transaction Process
TID_{req}	The request for TID
PID_{req}	Request for Patient Device Identity
{D}X	Data D symmetrically encrypted with shared key X
H(D)	The one-way hash function for Data D
I	Used to Identify the current session Key of Xi and Yi
K_{p_p}	The secrete key shared between Patients Device and Inter Health Institution LAN on

	registration
Success/Failed	The status of registration whether success or Failed
Yes/No	The status of Data transaction whether approved or rejected
Received	Information Receivable update status, which may include the description of data received

B. The Framework

The proposed Framework consists of two procedures, which are registration procedure and data transaction procedure. P_D and P_{DS} are required to register with their Network System before any data transaction can take place. P_D and P_{NO} create session key, X_i by running Diffie-Hellman Key Agreement protocol. P_D then sends registration details like Patient Names Phone Number and Patient Address encrypted with session key K_i to P_{NO}.

$$P_D \rightarrow P_{NO}: \{PNp, IDp, AIp\}\ K_1$$

A patient's details will be captured during first registration at a health center connected on the Inter Health Institution Local Area Network. This takes place ONLY on the InterHLAN network for accurate authentication, verification and integrity of patient data. The patient reports at the hospital or clinic reception and their initial bio data is captured by trained personnel, their health data is entered by a doctor and stored in a federated database as the first electronic health record.

During the Registration procedure, a Patient is required to set his/her Personal Identification Number, PIN_P for later access to his/her personal Medical data. The registration process also captures a patient's biometric finger print that shall be used on the InterHLAN in emergency cases or special cases such as a patient who forgets his Password. This procedure uses two factor authentication which is an vital principle for physical and mobile devices access control [21]. The two-factor authentication applies two methods to authenticate users to access the mobile based personal health information system. It uses the mobile device with the Application (something he/she has) and the Password (Something he/she knows only). Then the ID_P is calculated by hashing the PN_P and PIN_P.

$$ID_P = PN_P + H(PN_P\ PIN_P)$$

P_{NO} decrypts data with shared session key, K_1 to retrieve patient's health data. P_{NO} stores required data into their database. If registration procedure is successful, P_{NO} sends confirmation message to P_D to inform Patient. The confirmation message is encrypted with the session key K_1

$$P_{NO} \rightarrow P_D :\{ Success/Failed\}\ K_1$$

After registration process, Patient receives a mobile API access code from P_{NO}. The application contains symmetric key generation and mhealth software. After it has been installed successfully, a set of symmetric keys $X = \{X_1, X_2 \ldots Xn\}$ is generated and stored into P_D and sent to the P_{NO}. The P_D will also in the same process exchange a secrete key K_{P_P} with the Inter Health LAN authentication server during the registration process. Similarly, P_{DS} MUST go through similar registration process with the InterHLAN server to enable it to receive data from P_D. The P_{DS} generates a set of symmetric keys $Y = \{Y_1, Y_2 \ldots Y_n\}$ with the InterHLAN server and store them into the InterHLAN Database Server. The figure below shows the framework design. The various components of the framework are illustrated, and their interaction shown.

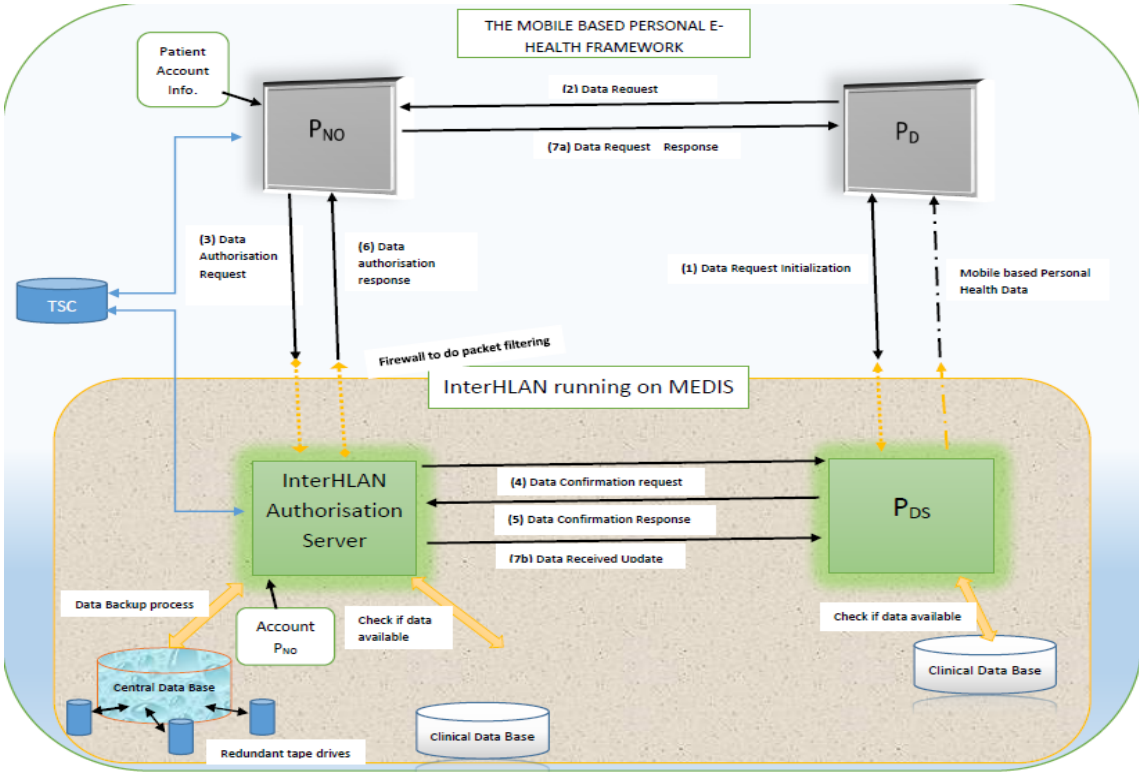


Fig. 2. The Framework

Phase 1 Data request Initialisation:
PD → PDS: *R, TIDreq, PDSIDreq*
PDS → PD: *{IDP, TID, PDSID} K2*

Phase 2 Data Request:
PD → PNO: *{IDP, PDSID, R, TID, ehr, DATE, Rand, H(IDPNO, R, TID, ehr, DATE, Rand), {R, DESC }K2 }Xi, i, IDP*
PNO → TSC: *H[{PDSID, PNOID, R, TID, ehr, DATE, Rand, H(PDSID, PDSID, R, TID, ehr, DATE, Rand)}, { R, DESC } K2 }Xi, i, IDP]*
TSC → PNO: *TimeStamp1*

Phase 3 Data Authorisation Request:
PNO → InterHLAN: *R, IDP, TID, ehr, DATE, {R, DESC} K2*

Phase 4 Data Confirmation Request:
InterHLAN → PDS: *{R, TID, ehr, DATE, {R, DESC} K2, Rand, H(R, TID, ehr, DATE {R, DESC} K2, Rand) H(KP_P) }Yi, i*

Phase 5 Data Confirmation Response
PDS → InterHLAN: *{Yes/No, Rand, H(KP_P), H(R, TID, ehr, DATE, {R, DESC}K2, Rand), {Yes/No, TID, ehr, DATE}K2}Yi+1*

Phase 6 Data Authorisation Response:
InterHLAN → TSC: *H({Yes/No, Rand, H(KP_P), H(R, TID, ehr, DATE, {R, DESC}K2, Rand), {Yes/No, TID, ehr, DATE}K2}Yi+1)*
TSC → PNO: TimeStamp2
InterHLAN → PNO: *Yes/No, TID, ehr, DATE, {Yes/No, TID, ehr, DATE} K2*

Phase 7 Data Request Response:
PNO → PD: *{Yes/No, Rand, H(KP_P), H(PDSID, IDNO, R, TID, ehr, DATE, Rand){Yes/No, TID, ehr, DATE}K2}Xi+1*
InterHLAN → PDS: *{Received, Rand, H(KP_P), H(R, TID, ehr, DATE, {R, DESC}K2, Rand)}Yi+1*

C. Emergency Cases Data flow:

Emergency cases are considered when a patient is unconscious, mentally unable to use the application on their mobile device, involved in a case with law enforcement agencies or has succumbed to fatal injury that may result in loss of life. The override to access a patients' data is made only possible with presentation of the patients' biometric finger print which was captured during the registration protocol using Gabor filter-based multiple enrolment fingerprint recognition [22] to protect patient data integrity and privacy.

Phase 1 Data request Initiation:
InterHLAN → P_{DS}: *R, TID_{req}, $P_{DS}ID_{req}$*
P_{DS} → InterHLAN: *{TID, ID_{NO}} K_3*

Phase 2 Data Request:
InterHLAN → P_{DS}: *{ Bio_P, ID_{NO}, R, TID, ehr, DATE, Rand, H(ID_{NO}, R, TID, ehr, DATE, Rand), {R, DESC }K_3 }Yi, i*

Phase 3 Data Request Response:
P_{DS} → InterHLAN: *{Yes/No, Rand, H($P_{DS}ID$, ID_{NO}, ID_P, Bio_P, R, TID, ehr, DATE, Rand){Yes/No, TID, ehr, DATE}K_3}Y_{i+1}*

V. STRENGTHS OF THE FRAMEWORK

The framework achieves a lot of strength in identity protection of the patient device (P_D) from the Patients federated data server (P_{DS}), identity protection from eavesdroppers or Man in the Middle attack, data transaction protection from eavesdroppers and transaction protection from trusted third party.

Patient's privacy protection is the most significant security achievement of the designed framework. The framework protects patient's identity by sending a random generated number (R) to patients federated data server (P_{DS}) when requesting the identity from the server InterHLAN. R symbolizes one-time patient's identity while the identity (TID) uniquely identifies Patients Device (P_D) to Federated Data Server (P_{DS}). This avoids revealing the real patient's identity (ID_P) to Patients Data Server (P_{DS}). The framework also provides transaction privacy from trusted third party (TTP) or Man in the Middle (MiM) attackers. The Data request that is sent from patient Device (PD) to patient's MNO consist the transaction details, which is *{R, DESC}K_2*. Note that, the data transaction details such as which data the patient wants or data from a number of visits is protected from both patient's MNO and Inter Health Institution Authentication server by encrypting it with the Patients Device (P_D) and Patients Data Server (P_{DS}) shared session key, K_2. Hence, only the corresponding Patients Data Server can decrypt and retrieve the data transaction details.
Besides that, both Data request message and Data confirmation response message are applied a hash function before sending it to Time Stamp Center (TSC). This prevents revealing of any data transaction details to TSC. To this end therefore, the proposed Framework comprehensively satisfies all privacy protection requirements for the patient.

The framework leverages on the pervasive computing concept to extend patient centered access control to resource limited environments. This concept is generic and does not meet electronic health systems requirements. The complements of certified access lists, message digests and secrete keys on the InterHLAN are integrated with pervasive computing to achieve the requirements of electronic health systems in resource limited environments.

VI. CONCLUSION

This research produced an access control framework for secure access to personal medical health records using mobile technology in resource constrained environments. Future work will involve building an actual system to implement the requirements gathered in this research and implementing it in the real world. Other information security domains such as business continuity and disaster recovery, application and system development security and physical security are outside the scope of this research, but future research can look into integrating them into this

framework. This research acknowledges that some patients may not be able to use mobile devices or computers. This does not negate their rights to privacy. There is need for future research on how to enable such patients to be able to manage access to their electronic health records.

REFERENCES

[1] P. Varkey, H., A., & K. E. Bennet (2008). "Innovation in health care: a primer". *American Journal of Medical Quality*, *11*(5), 382-388.

[2] U. C. Commission, "The Uganda Communications Commission Act," 17 September 2000. [Online]. Available: http://www.ucc.co.ug/ucaCap106LawsOfUganda.pdf. [Accessed 25 January 2017].

[3] R. Ssembatya (2014). "Designing an architecture for secure sharing of personal health records: a case of developing countries". PhD thesis, University of Cape Town, Cape Town, South Africa, 2014.

[4] M. I. Kim, & K. B. Johnson (2002). "Personal health records evaluation of functionality and utility". *Journal of the American Medical Informatics Association*, *9*(2), 171-180.

[5] R. Ssembatya, A. V. Kayem, (2015, March). "Secure and Efficient Mobile Personal Health Data Sharing in Resource Constrained Environments". *In Advanced Information Networking and Applications Workshops (WAINA), 2015 IEEE 17th International Conference*, 411-416.

[6] M. N. K. Boulos, S. Wheeler, C Tavares, & R. Jones (2011). "How smartphones are changing the face of mobile and participatory healthcare: an overview, with example from eCAALYX". *Biomedical engineering online*, *10*(1), 12.

[7] R. Ssembatya,, A. Kayem, & G. Marsden, (2013, January). "On the challenge of adopting standard EHR systems in developing countries". In *Proceedings of the 3rd ACM Symposium on Computing for Development* (p. 11).

[8] R. Ssembatya, S. Zawedde (2014). "Issues of Adoption: Can Health Services Designed for Developed Countries be adopted in Developing Countries?" In *INC*, 115-137.

[9] A. Dmitrienko, Z. Hadzic, H. Löhr, M. Winandy, & A. R. Sadeghi, (2011). "A Security Architecture for Accessing Health Records on Mobile Phones". In *HEALTHINF*, 87-96.

[10] J. D. Halamka, K. D. Mandl, & P. C. Tang (2008). Early experiences with personal health records. *Journal of the American Medical Informatics Association*, *15*(1), 1-7.

[11] I. Carpenter, M. B. Ram, G. P. Croft, & J. G. Williams (2007). "Medical records and record-keeping standards". *Clinical Medicine*, *7*(4), 196-319.

[12] T. Garrido, L. Jamieson, Y. Zhou, A. Wiesenthal, & L. Liang (2005). "Effect of electronic health records in ambulatory care: retrospective, serial, cross sectional study". *Bmj*, *198* (7491), 581.

[13] World Health Organisation (WHO). "Knowledge management strategy. Geneva", (2005) [Online]. Available: http://www.who.int/kms/about/strategy/kms strategy.pdf. [Accessed:15 August 2016].

[14] R. Sandhu, D. Ferraiolo and R. Kuhn (2001). "The NIST Model for Role-Based Access Control: Towards a Unified Standard". *ACM Transactions on Information and System Security (TISSEC)*.

[15] D. D. Bell, and L.J. La Padula (1976). "Secure Computer System: Unified Exposition and Multics Interpretation". No. MTR-2997-REV-1. MITRE CORP BEDFORD MA.

[16] A. Bourka, A. Kaliontzoglou, and D. Polemi (2003). "PKI-based Security of Electronic Healthcare Documents". *Proc SSGRR*.

[17] I. Mavridis, C. Georgiadis, G. Pangalos, and M. Khair (2001). "Access Control Based on Attribute Certificates for Medical Intranet Applications". *JMIR***3**(1), 9.

[18] D.K. Vawdrey, T.L. Sundelin, K.E. Seamons and C.D. Knutson (2003). "Trust Negotiation for Authentication and Authorization in Healthcare Information Systems". *Proc Annual International Conference of IEEE,* 2:1406-1409

[19] S. Snezana (2007). "Implementing Security in a Distributed Web-Based EHCR". *International Journal of Medical Informatics,* 491-496.

[20] S. Avancha, A. Baxi, & D. Kotz, (2012). "Privacy in mobile technology for personal healthcare". *ACM Computing Surveys (CSUR)*, *45*(1), 3.

[21] R. R. Panko (2004) "Corporate Computer and Network Security", *Prentice Hall*, Upper Saddle River, New Jersey.

[22] F. Kaggwa, J. Ngubiri. (2016, April). "Gabor Filter-based Multiple Enrollment Fingerprint Recognition". *International Journal of Computer Applications, 139*(7), 32-38.

Geo and Graph Analytics for Dynamic Cellular Transactions Insights, Improving Quality and Business Decisions: "Quality X Map"

Dahj Muwawa Jean Nestor
Department of Electrical & Electronics Engineering
University of South Africa
Pretoria, South Africa
dahjmuwawa@gmail.com

Kingsley A. Ogudo
Department of Electrical & Electronics Engineering
University of Johannesburg
Johannesburg, South Africa
kingsleyo@uj.ac.za

***Abstract*— Everyday, millions of users on their phone generate huge amount of traffic; intelligent industrial machines and IoT devices also generate monumental traffic, both affecting Network performance due to the shared medium used. With the large amount of human and machine generated data, Communications Service Providers (CSPs) face the challenge of finding values and addressing Quality of Service (QoS) and Quality of Experience (QoE) in a dynamic way. The advance of Data and Predictive Analytics methods makes it not easier but possible to dive deep into Network transactions to build intelligent insights, helping with business decisions. In this paper, we use Geo-Analytics and Intelligent Localization algorithms, combined with Graph Processing to provide a dynamic insight of Network data, to sustain business decisions and improve users and devices' QoE.**

Keywords—Geo-Analytics, Graph Analytics, Cellular Transactions Insight, Quality of Experience (QoE), Business Decisions, Internet of Things (IoT)

I. INTRODUCTION

As Data Analytics is taking its toll across all the sectors of businesses, so it is in the Telecommunications environment. With massive generation of data records by subscribers, Communication Service providers (CSPs) face challenges in transforming the data to actionable values for decision making to improve QoS and QoE. Hence, big investments on Data Analytics, Customer Experience Management (CEM) and Service Quality Management (SQM). It is therefore, not objective to only take advantage of the fast processing capabilities of the Data Analytics systems, but to also provide visualization in an understanding manner, supporting business decisions. Data visualization, with the advancement of computing power, has taken a different perspective [1]. Subscribers are constantly using their devices, meaning the size of network traffic or transactions keep growing, accelerating the need for fast, efficient and user-friendly data analytics systems to improve QoS and QoE, enabling an adequate decision-making process. The diversion between QoS and QoE resides in a single focal point, the perception parameter. While QoS focuses on "Service" delivery index, QoE focuses on "Experience" by the user. The two concepts are the baseline of SQM and CEM respectively. Estimating the number of transactions that can be generated in a period by telephones users, we introduce an effective, realistic and simple way to visualize network transactions information using Geographic Information System (GIS) and graph processing algorithm. The objective of the paper is to go beyond the popular data visualization schemes, such as charts (pies, bars, lines), maps and tables for data insights, to introduce the "where" parameter in Geo-Analytics and exploring the benefits of Graph Processing. Graph algorithms have been in use for a long time in the history of computer science and Data Analytics [2]. Fig. 1 illustrates the concept of the study which explores the advantage of spatial Analytics, Geo databases functionalities, flexibility and agility of graph processing.

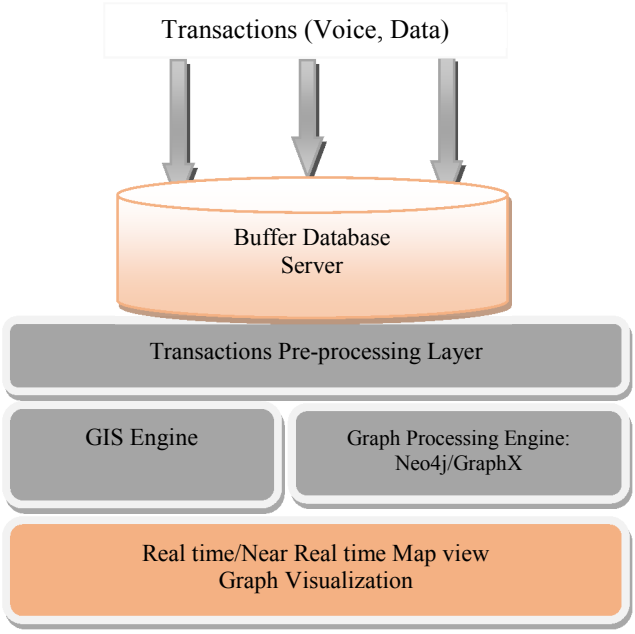


Fig. 1. Study Concept and Model

In this paper, we use Geo Analytics and Graph Processing model to analyze user data trafic by exploring our Quality Exploration Map model (QXM) which rest on the application layer (visualization). The lower layers are subject to customization depending on the needs, ranging from Structured to unstructured data.

II. Study Review

Traditional and popular data visualization models have been mostly evolving around tables, charts, graphs, lists to spice up business processes; however, the liaison between data and its location is not addressed using such methods of visualization. Several researches have pinpointed the benefits of Geo-Analytics and visual interaction as an effective data visualization method. Angi Voss et al. explores sales Analysis of retail companies using commonGIS and OLAP systems [3]. Geo-Analytics has also been a key model in the transportation industry, with research such as E. Packer & al. [4], which emphasizes on visual analytics for spatial clustering. The study uses Heat maps to efficiently display the circulation of public transports cars in different levels. Visual queries and geographical representation provide a flexible access to explore relationship between different attributes or components of the dataset. R. Shadom and C. Weaver explores the strength of visual Analysis and intra and inter component relationships using hypergraph query methodology [5]. GIS (Geographical Information Systems) applications are used in many industries. Data Analytics using graphs and locations as differentiator points facilitates multidimension exploration of attributes or components relationships [6].

The use of Graph Processing in the past has been remarkable in areas such as Social Networks with big companies such as Facebook, LinkedIn, Twitter adopting the technology and modeling of diverse web traffic [7]. Many researches also are being conducted around graph processing, with the development of technologies such as Spark GraphX [8].

Jakob Smedegaard Andersen and Olaf Zukunft used GraphX to evaluate the scaling of Graph-Algorithms for Big Data [9], in which they analyze social media graphs using semi-clustering method. The study explores relationship and interactions between users, the addition of friends, channel subscription, message exchanges etc. Relationships have certain weight depending on the interactions and messages exchanges. Vertices connote users and Edges connote relationships between users. The objective of Jakob and Olaf's study of social behavior was to define a certain number of user groups defining a certain criterion such as strong relationships with another category of user group. The relationship equation is given by the equation (1), given a cluster of users c, a set of relationships or properties S:

$$Sc = \frac{I_C - f_B B_C}{V_C(V_C - 1)/2} \qquad (1)$$

Where with Ic the sum of the weights of all relationships inside the cluster c, Bc the sum of weights of all relationships that go outside the cluster, fB a factor to Bc, which is assumed to be in the interval [0, 1] and Vc is the number of vertices, here representing users in the cluster. The output of a semi-clustering is illustrated in Fig. 2.

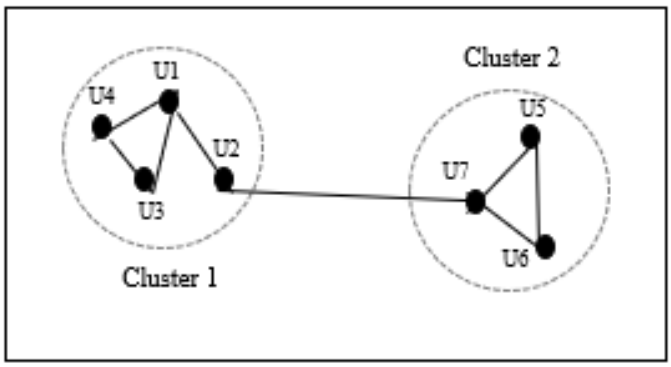


Fig. 2. Illustration of a semi-clustering Graph Algorithm Output

The semi-clustering of graph modeling has been studied by G. Malewicz & al. in their Pregel model [10]. However, the model is undirected. Graph processing used in GraphX or Neo4j are mostly directed. Using the same approach, transactions from Network data users are analyzed based on a set of criteria to provide the Communication Service Providers, a ground to improve Quality.

On the other side, the use of Geo Location has been widespread to give a different insight of spatial data; such examples include the use of heatmaps, density maps, 3D Maps in different areas such as Network Cells planning and Optimization, mobile phones Geolocation and many more. Exploiting the power of Geo Analytics facilitates the understanding of traffic distribution. Michelle Angelico & al. used Smart Data Localization, based on Geo Analytics of environmental and territorial data to improve quality of life by promoting the development of Business Support platforms [11]. Jianghua Zhao & al. go even further using GIS system to predict traffic noise, segmenting the roads which at sensitive points, contributes with the highest noise level [12].

The Continuous QoS and QoE song: Communication Service Providers are always taking quality as a serious topic. It is a song that will continuously be played because Network Infrastructures are continuously changing. Good network quality gives competitive edge to CSPs [13]. QoS and QoE acceptable threshold is a parameter that needs to be determined by the CSP, following specific standards and classes [14]. The concept of Intelligent localization algorithm relates to the use geospatial information [15], to pin-point service problems such as traffic load, internet data usage. The communication environment applies geographic data in diverse operations including Line of Sight Analysis for Microwave Transmission signals, users and event geolocations [16]. The model pushes the use of Geo Intelligence and Graph processing towards Analytics and Customer Experience

III. Contribution of the Study

The use case aims to provide an efficient visualization of network traffic displaying, at the same time the quality indicators. The use case model or approach provides the CSPs with a visualization that facilitates the detection of

abnormalities in the Network. Using this method, CSPs not only saves on CAPEX investment, they leverage on the openness and advancement of technologies such as Big Data, Prediction Analysis, Geo Analytics etc. Adopting a graph and visualizations as approached in this paper also gives CSPs a multi-department use case platform for decision making. The area of Business Intelligence and Data Visualization is an evolving area and always leave gaps for improvements.

IV. Methodology and Design

When referring to Analytics a certain number of processes and models are overlooked. From data collection to data visualization, complex methods and algorithms can be considered. However, in this paper, data preprocessing is considered done as part of the Back-End functionalities. Fig. 3 shows the steps used for this research starting directly with Exploratory Data Analysis.

- Exploratory Data Analysis: to understand the dataset and the meaning of variables and exploration of indicators. Exploratory Data Analysis also allows us to establish pre-relationship between different attributes of the dataset. For example, establishing a link between a service application and its required key performance index.
- Data Query Aggregation: to create and facilitate the modeling of vertices and edges of the dataset. This is the fundamental use case step to support decisions in the model. The queries are built based on the computation components. If the component is the service, in the case of computing QoS, data is aggregated around the services, assigning weights to the edges or links between the different services.
- Data Visualization: data presentation layer, mostly based on a Map or/and graphical processing view. For this paper, Google Map, Spark GraphX and Neo4j are used with R to illustrates the model. This is the step to transform and present the processed data into an understandable measurement.

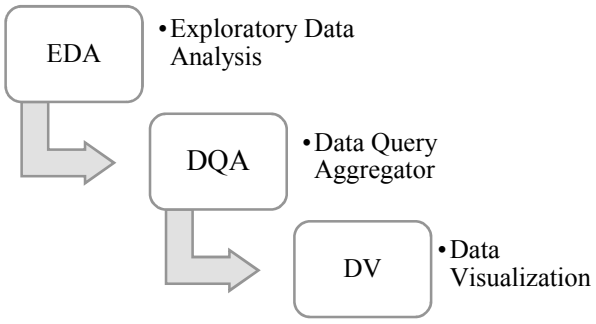


Fig. 3. Design Process used in this Paper

A. Quality X Map Model

The visualization Model is built on Graph processing approach overlaid with GIS view. "X" refers to the indicators or aggregation of indicators which can be used to visualize the quality level, also defined as a factor of Exploration. X can be indicator, score, aggregator, count, mark depending on the applications. The model combines GIS and Graphical views. The model explores node to link blueprint to illustrate relationships between attributes of the dataset. In a transactional database, it can be requested to find streaming "Netflix" service usage across the country and which gateway server cater most of the streaming traffic. QXM model facilitates the computational (using graph processing) and visualization (geographical view).

Given a Structured Dataset D with variables k, u, v and c, the dataset can be represented as:

$$D(u,v,c) \rightarrow k^i|u^i|v^i|c^i \qquad (2)$$

Where k and u are the variable keys of the dataset, v is the variable Indicators (KPIs, Calculated score, count) and c is the spatial variable. Explicitly, the dataset can be shown as a transaction model:

TABLE1. TRANSACTONAL DATASET MODEL

k	u	v	c
k^1	u^1	v^1	c^1
k^2	u^2	v^2	c^2
.	.	.	.
.	.	.	.
k^i	u^i	v^i	c^i

The Quality X Map allows us to represent the transactions in the dataset in holistic manner, giving at glance the geographical points of interest as illustrated in Fig. 4. The Figure shows in real time, near-real time or historical manner, the transactions analysis. On the Graph processing, k and u are vertices of the graphs systems and v represents the edges of the graphs as shown in Fig. 5. The model can be used for streaming applications where the edges are directed as files are sent. To represent data on the GIS, the dataset must have spatial information, either as coordinates or administrative regions. Therefore, for Geo Intelligence, c must be present.

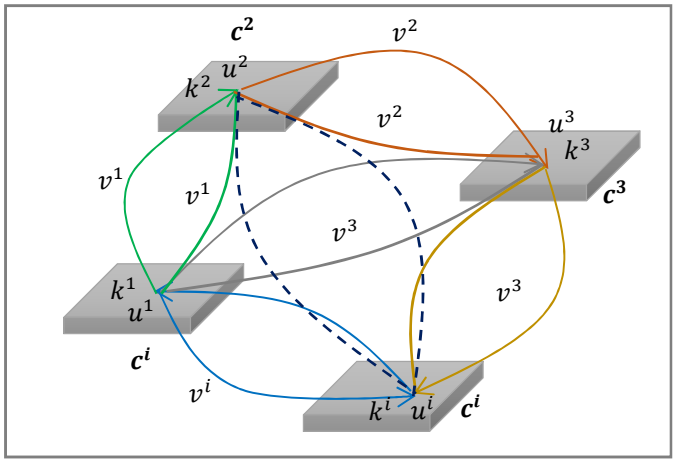


Fig. 4. Quality X Map model for GIS view

v^i is the Quality variable that represents a relationship or interaction between two dataset keys k, v by geographical region or coordinates c^i. Typical use case, number of data sessions between Gauteng and Western Cape servers; active subscribers YouTube streaming sessions towards a specific Gateway server.

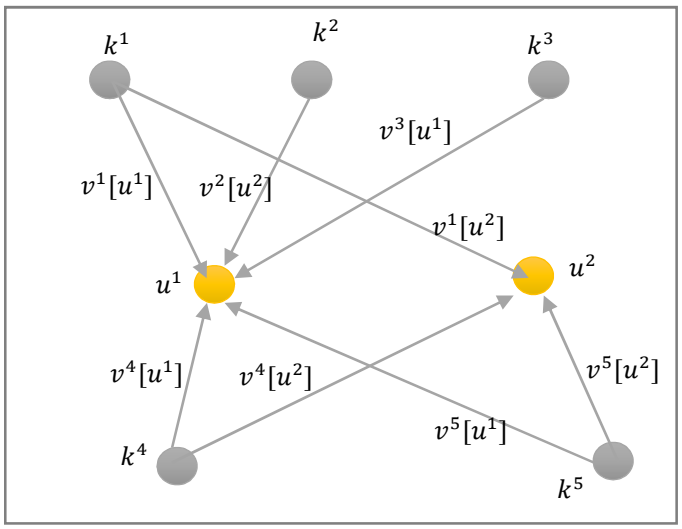


Fig. 5. Quality X Map Model Graphical view

$v[u]$ is the quality index towards the node u and vertices whose weight determines the value of X. And k is the vertices and start nodes for transactions. Example, analyzing the transactions towards a server to check for possible congested links and failures. In the Graphical view, the parameter c is not required unless it is taken as vertices to enable a graphical view aggregated on c.

B. Quality X Map for Mobile Data Transactions

In this section, we use public Telecoms data to apply the Quality Map model in a real dataset. We use quantitative QoS (Quality of Service) [16]. Data from an Internet Service Provider is analyzed to provide the overall performance of Services provided to customers. The summary of the data is shown in table 2.

TABLE 2. MOBILE DATA DATASET SUMMARY FOR ANALYSIS

Quality variable	**Unit**	**Data type**	**Role**
AccessID	None	Varchar	Vertex
InternetGateway	None	Varchar	Vertex
ProtocolApp	None	Varchar	Vertex
ActiveDataSessions	None	Bigint	Edge
DataUSage	Gbytes	Bigint	Edge
RtxPacket	Gbytes	Bigint	Edge
packetloss	%	Decimal	Edge
Ul_latency	msec	Decimal	Edge
Dl_latency	msec	Decimal	Edge
Dlthroughput	Mbits/sec	Decimal	Edge
Ulthroughput	Mbits/sec	Decimal	Edge
Region	None	Varchar	Can be used as Vertex.
Latitude	None	Decimal	Can be used as Vertex
Longitude	None	Decimal	Can be used as Vertex

Region (administrative), Latitude and Longitude (Geographical Coordinates) are spatial variables which can also be used as vertices for regional aggregated values.
Using GraphX, the below use cases are evaluated using Quality Map model.

1) Real time Network Transactions and Congestion Analysis using Quality Map Model.

CSPs and ISPs (Internet Service Providers) have to keep a close look on Server's load to avoid possible server crashes and overload. While many mechanisms and algorithms of congestion control are being studied [17], one of the challenges is to oversee possible congestion in links. Figure 6 illustrates the Geo representation of the Quality X Map Model for traffic links analysis. Based on table 2, the ISP has 2 Gateway servers for national internet connection. The model provides the simultaneous active data sessions per server, giving the opportunity to view any other Indicator related to traffic Analysis. Fig. 6 dynamically shows the amount of sessions originating from different access sites towards the gateway servers. Due to load balancing, Access sites can connect to both servers. Fig. 7 shows the same use case based on Graph processing using iGraph with R. Representing the same data on a chart could be difficult to read (~ 100 access points).

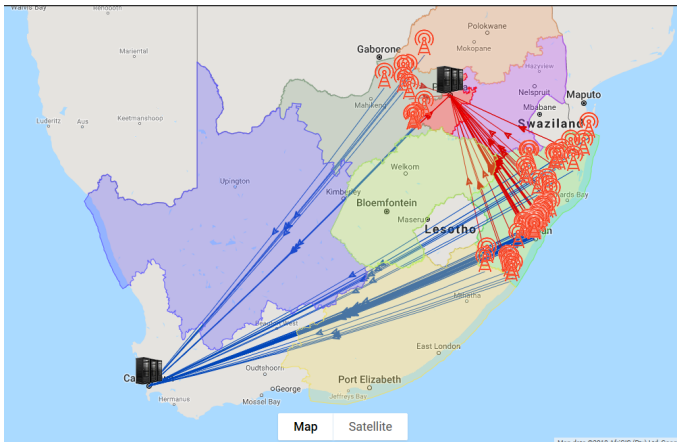


Fig. 6. Real time Geo Analytics of Server transactions for traffic load

The two servers of the CSP are shown in Fig. 6 with one big gateway server in Pretoria and the other one in Cape Town. The tower represents the access sites from the users. From the Map, provide the ISP with the ability to pin-point the server having excessive traffic loads. The Model displays the traffic flows as it happens, in near real time.

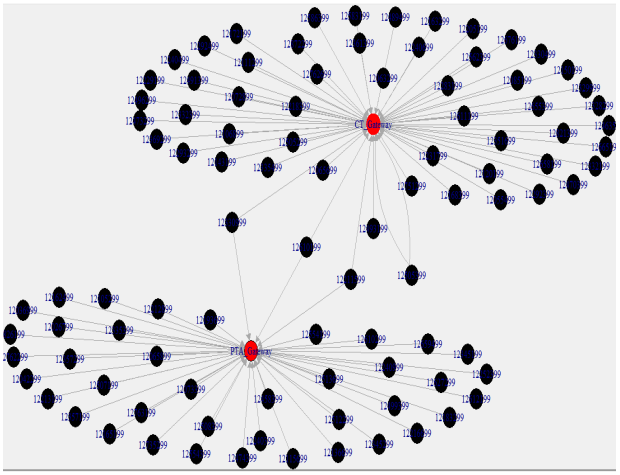


Fig. 7. Simultaneous Active Data sessions per Gateway Server

Fig. 7 displays the same analysis in a Graphical way, reducing the possibility to overcrowd the data space and ease recommendation processes.

2) NetFlix Traffic Analysis using Quality X Map Model

The objective is to use the model to provide NetFlix (video-streaming) traffic Analysis per Region. Applications for which may be linked to sales, marketing and Operations.

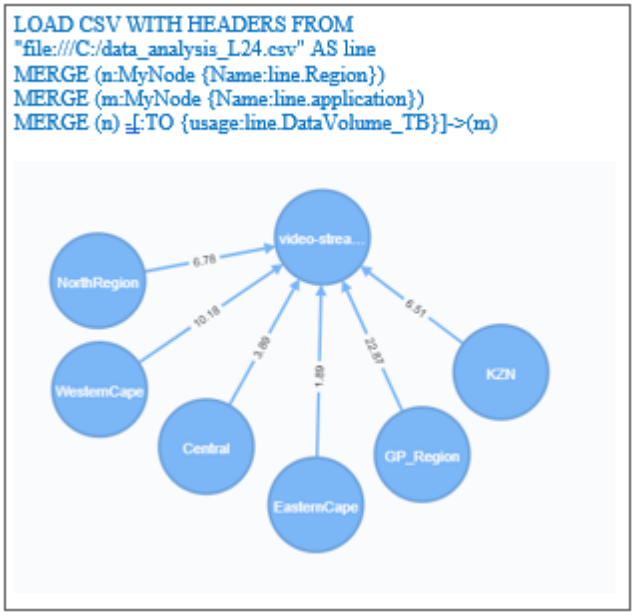


Fig. 8. NetFlix Traffic Analysis Per Region

Fig. 8 displays the Graph Model of Analyzing traffic distribution of specific traffic using Neo4j platform. The edges display the total number of data used in Terabytes. The edges represent the interaction between different nodes or vertices. Any other metric could ne used in place of data volume.

3) Customer Call Analysis using Quality X Map Model

The use case uses the Quality X Map to plot the user call activities in a dynamic way. The application of which could be linked to customer experience and fraud detection. Fig. 9 illustrates the Analysis of group call trace.

The target subscriber is ***17 (Subscriber anonymized for privacy purposes). The analysis exposes the call transactions from and towards the same group of people with user ****17 as reference. The users are vertices and the number of calls is edges, showing the interactions between vertices (subscribers). ****17 made 5 calls to ***43 who made 2 calls to ****80. Subscriber ****80 then called ****17 back 10 times.

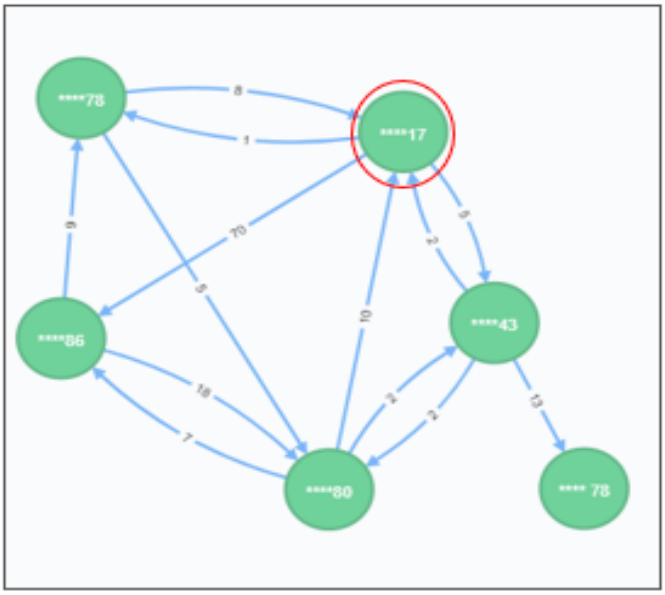


Fig. 9. Network Graph of Customer Call Analysis

V. Conclusion

Information visualization is a capital part of Data Analytics process, because it is through visualized reports and dashboards that business decisions are taken. Overlooking beyond the traditional forms of data presentations, the Quality X Map model as approached in this paper allows an instantaneous view of Network status and performance in an easy and understanding way for all. Graph processing is becoming more and more an efficient way to analyze and visualize data. Our study combines the benefits of Graph processing and GIS to provide a different business Insight. The model can be adjusted to fit any other area beside Telecoms. The Quality X Map provides benefits such as flexible data visualization in which every network transaction can be represented as a graph or as GIS if it contains spatial information, very easy to expend the data by adding vertices and easy to use model for Real time monitoring.

VI. Future Work

With the increase in cellular usage, a certain level of dynamism should be adopted in handling transaction data [18]. Geo Analytics in combination with Graph processing as described in this paper, provides accessibility to critical network information at any location [19]. Telecoms transactions are of a big magnitude because of the number of

events executed by Mobile devices, therefore, require Geo-Analytics and Graph Processing for smart reporting, prediction and recommendation capabilities [20]. Looking at Cellular Network planning, the use of Geo Analytics can be of capital use [21], [22].

The model has been applied to a small set of datasets, with less emphasis on data pre-processing. The future work will focus on the analysis of roaming traffic, which we believe would be very suitable for the Quality X Map model of Visualization. We will include large network dataset, taking advantage of Big data processing mechanisms such as Hadoop, Spark in-memory computing [24] to speed up the processing of real time and batch big data. That also means that the computing power of the current server will be improved (RAM, Storages) to allow the processing of large datasets. We also look at using some clustered algorithms to enhance the model and the span of application. We have used two different Graph visualization platforms to demonstrate the model (Neo4j and iGraph with R); However, as data increases, we will use the Spark GraphX to handle more data. The future work will include the analysis of various domains of network as shown in Fig. 10, including network performance scores, Quality Score Map.

Fig. 10. Future work' s scope for Quality X Map Model

REFERENCES

[1] A. Azzalini and B. Scarpa, "New problems and new opportunities", in *Data Analysis & Data Mining, and Introduction*, Ed. By Oxford University Press, 2012, pp 2-8.

[2] R. Diestel, Graph Theory, 4th ed. Heidelberg: Springer, 2010.

[3] A. Voss, V. Hernandez, H. Voss and S. Scheider, "Interactive visual exploration of multidimensional data: requirements for CommonGIS with OLAP," *Proceedings. 15th International Workshop on Database and Expert Systems Applications, 2004.*, Zaragoza, 2004, pp. 883-887.

[4] E. Packer, P. Bak, M. Nikkilä, V. Polishchuk and H. J. Ship, "Visual Analytics for Spatial Clustering: Using a Heuristic Approach for Guided Exploration," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2179-2188, Dec. 2013.

[5] R. Shadoan and C. Weaver, "Visual Analysis of Higher-Order Conjunctive Relationships in Multidimensional Data Using a Hypergraph Query System," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2070-2079, Dec. 2013.

[6] N. Henry and J.-D. Fekete. MatrixExplorer: a dual-representation system to explore social networks. *IEEE Transactions on Visualization and Computer Graphics,* 12(5):677–684, September/October 2006

[7] A. Silberstein, A. Machanavajjhala, and R. Ramakrishnan, "Feed following: The big data challenge in social applications," in Databases and Social Networks, ser. DBSocial '11. New York, NY, USA: ACM, 2011, pp. 1–6. [Online]. Available: http://doi.acm.org/10.1145/1996413.1996414

[8] J.E. Gonzalez, R.S. Xin, A. Dave, D. Crankshaw, M.J. Franklin and I. Stoica, "GraphX: graph processing in a distributed dataflow framework", OSDI'14 Proceedings of the 11th USENIX conference on Operating Systems Design and Implementation, Broomfield, pp: 599-613, 2014.

[9] J.K. Andersen and O. Zunkuft, "Evaluating the Scale of Graph-Algorithms for Big Data using GraphX", IEEE 2nd International Conference on Open and Big Data, 2016.

[10] G. Malewicz, M. H. Austern, A. J. Bik, J. C. Dehnert, I. Horn, N. Leiser, and G. Czajkowski, "Pregel: A system for large-scale graph processing," in Proc. of the 2010 ACM SIGMOD Intl. Conf. on Management of Data, ser. SIGMOD'10. New York, NY, USA: ACM, 2010, pp. 135–146.

[11] M. Angelaccio, B. Buttarazi, A. Basili and W. Liguori, "Using Geo Business Intelligence to Improve Quality of Life", IEEE International, 2012.

[12] J. Zhao, Q. Qin, C. Xie, J. Wang, and Q. Meng, "AN EFFICIENT METHOD OF PREDICTING TRAFFIC NOISE USING GIS", Institute of Remote Sensing and GIS, Peking University 10087 1, Beijing, China, 2013 IEEE.

[13] Power, J. D., and Associates (2003), "Wireless network quality assessment study™". Retrieved November 11 th, 2004, from http://www.jdpower.com/studies/pressrelease.asp?StudyDD=891.

[14] 3GPP TS 23.107 version 5.4.0 Release 5, Universal Mobile Telecommunications System (UMTS); Quality of Service (QoS) concept and architecture, Technical Specification.

[15] J. Pan, H. Wang, and F. Gan, "Analysis and Designing of GIS-Based Graph and Image Database System of Petroleum Exploration,"Earth Science, vol. 27(1), 2002, pp. 59-62

[16] Y. Cao and G. Yang, "The application of geospatial data technique based on GIS on oil field exploration & development graphic and documents' management database," *2011 19th International Conference on Geoinformatics*, Shanghai, 2011, pp. 1-4.

[17] J. Lempiäinen, M. Manninen, Radio Interface System Planning for GSM/GPRS/UMTS, Kluwer Academic Publisher, 2001.

[18] R. Rodríguez, D. Fernández, H. Montes, S. Hierrezuelo, G. Gómez, "Quality of Service Mechansms" in *End-to-End Quality of Service over Cellular Networks*, Ed. John Wiley & Sons, 2005, pp 103-137.

[19] S. Fang, C.H. Foh, K.MM.Aung, "Prompt Congestion Reaction Scheme for Data Center Network using Multiple Congestion Points", IEEE ICC Next Generation Networking Sympsium, pp. 2679-2673, 2012

[20] Brockmann, T., Stieglitz, S., Kmieciak, J., and Diederich, S., 2012. "User Acceptance of Mobile Business Intelligence Services". 2012 15th International Conference on Network Based Information Systems, vol. 23, pp. 861-866.

[21] P. Gray, "Business Intelligence: a new name or the Future of DSS?" in *DSS in the Uncertainty of the Internet Age*, Eds Bui T., Sroka H., Stanek S. Goluchowski J., Publisher of the Karol Adamiecki University of Economics, Katowice, 2011.

[22] F. Facchini, R. Giuliano, F. Mazzenga, "UWB Detect and Avoid Procedure for WiMAX victims", IET Commun., Vol.3, Is.2, Feb. 2009, p.268-278.

[23] R. Giuliano, F. Mazzenga, "Exponential Effective SINR Approximations for OFDM/OFDMA-based Cellular System Planning", IEEE Trans. Wir. Com. (TWC2009), Vol.8, No.9, Sep. 2009, p.4434-4439.

[24] R. Giuliano, P. Loreti, F. Mazzenga, G. Santella, "'Fractional Frequency Reuse Planning for WiMAX over Frequency Selective Channels", IEEE Int. Wirel. Comm. Mob. Comp. Conf. (IWCMC 2008), Chania Greece, Aug. 2008, p.666-671

[25] J. Betser, D. Belanger, "Architecting the Entreprise via Big Data Analytics" in *Big Data and Business Analytics*, Ed. by Jay Liebowitz, CRC Press: Taylor & Francis Group, 2013, pp 2-50.

Emulating Software Defined Network Using Mininet and OpenDaylight Controller Hosted on Amazon Web Services Cloud Platform to Demonstrate a Realistic Programmable Network.

Lindinkosi L. Zulu
College of Science, Engineering and Technology
University of South Africa
Florida, Johannesburg, South Africa
61366935@mylife.unisa.ac.za

Kingsley A. Ogudo
Department of Electrical and Electronic Engineering Technology
University of Johannesburg
Johannesburg, South Africa
kingsleyo@uj.ac.za

Patrice O. Umenne
Department of Electrical and Mining Engineering
University of South Africa
Florida, Johannesburg, South Africa
umennpo@unisa.ac.za

***Abstract*— In this paper, a Software Defined Network was created in Mininet using python script. An external interface was added in the form of an OpenDaylight controller to enable communication with the network outside of Mininet. The OpenDaylight controller was hosted on the Amazon Web Services elastic computing node. This controller is used as a control plane device for the switch within Mininet. The OpenDaylight controller was able to create the flows to facilitate communication between the hosts in Mininet and the webserver in the real-life network. In order to test the network, a real life network in the form of a webserver hosted on the Emulated Virtual Environment – Next Generation (EVE-NG) software was connected to Mininet.**

Keywords—SDN; Mininet; OpenDaylight; Amazon AWS, Cloud networking and Cloud computing

I. INTRODUCTION

Traditional networking is currently very decentralized with each network device having its own control plane. It requires individual manual configuration of each device on the network if there are changes to be implemented. The hardware and the software of the traditional networking architecture are proprietary and specifically designed to work together. This current setup makes it difficult for the network to be flexible and scalable to meet the high demand of modern applications and requirements. Software Defined networking (SDN) was developed to address these challenges the current network model is failing to address and OpenFlow was developed as the first standard communications interface defined between the control and forwarding planes of an SDN architecture [1].

Open Networking Foundation (ONF) defines Software Defined Networking as the physical separation of the network control plane from the forwarding plane, whereby the control plane controls several devices externally. It is an architecture that decouples the network control and forwarding functions. This allows the network to be dynamic, adaptable, cost-effective, software programable and easily manageable [2]. The advent of cloud computing allows the control plane to be logically centralized and distributed in cloud platforms.

This paper begins by looking at the literature covering Software Defined Networking, Mininet, OpenDaylight controller and cloud networking. The methodology section follows which describes the network used and its components. On the results section, the test results such as throughput are documented. The conclusion summarizes key points and is followed by the references.

II. BACKGROUND STUDY

The control plane is the centrally located control unit called SDN controller acting as the Network Operating Systems (NOS). The data plane resides inside the network core devices and is only responsible for forwarding data packets controlled by the central SDN controller. These separated planes use protocols and an Application Programmable Interface (API) to communicate [3].

OpenFlow is one of the protocols used by Software defined networks and was started by Stanford University in 2008 [4]. Different companies came together in 2011 and formed Open Networking Foundation (ONF) to further develop OpenFlow and Software Defined Networking [5]. With the separation of the control and data planes, the data plane only performs the data packet forwarding action and it resides in the network device. The control plane is logically positioned on top of the data plane and acts as the brains of the network [6]

Software Defined networks makes it possible to consolidate in one place complex software used to configure and control several devices making the process less expensive. A centralized controller gives a benefit of having a view of the

network, which then enables it to make decisions on how data planes must move the traffic [7-10].

SDN makes it possible to dynamically provision the network. It improves network resources utilization and simplifies traffic engineering [11]. It makes it possible to use external applications to program the network. Communication between the devices in SDN uses open interfaces making it to be vendor neutral [12]. To test Software Defined Networks, an emulator called Mininet is amongst the popularly used tools [13].

A. Mininet

Mininet is the container-based emulator [14]. It allows the running of unmodified code interactively on virtual hardware on a regular computer. It provides convenience and realism at low cost compared to running on a hardware. Programs run on emulators require none or minimum modification when applied to real live networks [15]. Mininet runs unmodified code of network applications in lightweight Linux containers to achieve its scalability and accuracy.

Mininet supports OpenFlow-based Software-defined Networking (SDN). It provides a flexible and cost-efficient experimental platform to develop, test, and evaluate OpenFlow applications. In Mininet, the processes of the virtual hosts and their application processes run inside the container. This allows them to have an independent view of system resources but still share the kernel with other containers [16].

Mininet supports five built-in network topologies. These built-in topologies are Minimal, Single, Linear, Tree and Reversed. Network topologies in Mininet can be modified using the command-line interface (CLI) [17].

The study by Keti and Askar in [18] highlights Mininet's characteristics as being flexible, applicability, interactivity, scalability, realistic and share-able. This is because in Mininet, new topologies and new features can be set in software using programming languages and common operating systems. Networks emulated in Mininet are usable with real life networks based on hardware without the need to make changes in source codes. To manage and run the simulated network in Mininet occurs in real time as it happens in a real-life network. Mininet can be scaled to large networks with hundreds or thousands of nodes. Networks implemented on Mininet can be easily shared as it is share-able [19].

Software Defined Networking switches, hosts, controllers and links can be created by typing commands through Mininet's command line interface. The command line interface (CLI) in Mininet supports most Linux commands. The most commonly used commands are-: nodes: which lists all created nodes, dump: which displays the information about the network and created nodes, net: which shows how network elements are connected to each other. The CLI also support the day-to-day troubleshooting commands used in computer networking. These commands include "pingall", which output the results of the connectivity test among all nodes. "Ifconfig" is also supported which displays the internet protocol (IP) information of the node. "Iperf" is also supported which is a tool used to test network performance. It uses a client/server model, where traffic is initiated from the client and traverses the network to the server. Iperf creates data test streams supported by the network with a time-stamp and report the amount of data transferred and the throughput measured. Iperf supports two types of transport protocols: Transmission Control Protocol (TCP) and User Datagram Protocol (UDP). Many applications like File Transfer Program (FTP), Simple Mail Transfer Protocol (SMTP) and Hypertext Transfer Protocol (HTTP) use TCP as the transport protocol. Using TCP mode, Iperf tests the maximum TCP bandwidth at the transport layer. In UDP mode, Iperf tests the jitter, packet loss and bandwidth. UDP mode is ideally for testing quality of service for applications like voice and video streaming.

To see the list of all available commands, one can use help command, which is also supported in Mininet [20].

Mininet Python API can also be used to create custom network topologies [21]. Python is an interpreted, interactive, object-oriented programming language. It provides high-level data structures. Python is modular by nature. The kernel is very small and can be extended by importing extension modules. A python program is compiled automatically by the interpreter and can be installed in any computer running any operating system [22].

B. OpenDaylight Controller

OpenDaylight (ODL) controller is the Software Defined Networking controller used in this project. It is based on Services Abstraction Layer (SAL), which allows it to support other protocols and not only OpenFlow. It is implemented in Java and can be deployed in any system supporting Java. OpenDaylight controller was developed by the OpenDaylight consortium in 2013. OpenDaylight project is supported by Cisco, Juniper, VMWare and many other vendors and companies operating in the networking environment. This support by many organizations enables OpenDaylight to be vendor neutral [23]

Fig 1 shows the OpenDaylight controller architecture. The controller uses Application Programable Interfaces (API) like Representational State Transfer (REST) technology to communicate with the Network Applications orchestrations and services layer. This can include OpenStack Neutron, Virtual Tenants Network (VTN) coordinator [24-25]. The controller layer itself run several services which includes service abstraction layer (SAL), OpenStack service, base network service and many more. To communicate with data plane elements, the controller uses southbound interfaces and protocol plugins such as OpenFlow, the Open vSwitch Database Management Protocol (OVSDB) [26], the Network Configuration Protocol (NETCONF) [27] and many more [28].

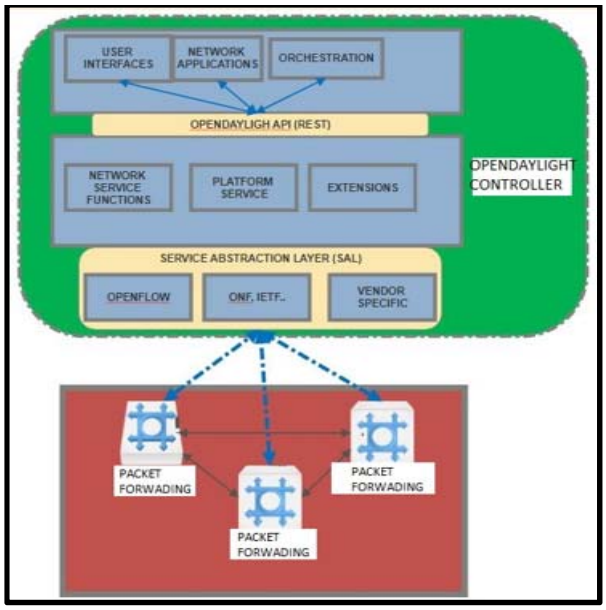


Fig. 1 OpenDaylight controller architecture

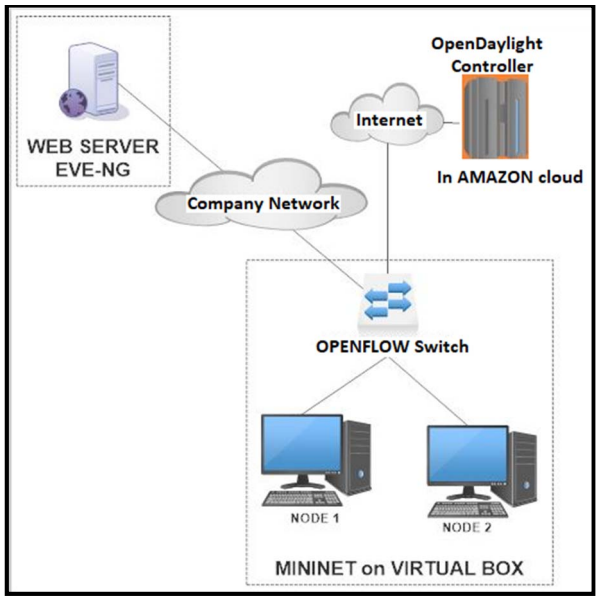


Fig. 2. Logical network used on this paper

C. *Cloud Computing*

The centralization of control plain and other SDN features has made cloud computing to be the essential part for most companies. It is one of the fastest emerging business for Internet Service Providers (ISP) [29]. A cloud can be described as a large-scale environment that can consists of many physical hosts and virtual machines (VMs). Each host in the cloud environment can serve multiple virtual machines. Services on the cloud network are provided on-demand bases. Virtual machines in a physical host can be dynamically provisioned as per the need. They can be added, removed or even migrated dynamically [30].

Amazon Web Services (AWS) is the leading cloud provider that offers computing, storage, and content delivery platforms. Amazon cloud services includes Elastic Compute Cloud (EC2) [31] and Simple Storage Service (S3), with "CloudFront", the Content Delivery Network (CDN). Amazon through AWS offers a large set of computing resources, such as storing and processing where capacities can be split, assigned dynamically as per customer's needs. Companies like Netflix and Dropbox are among the companies that uses Amazon Web Services (AWS) [32].

III. Methodology (Network Design)

The network used in this paper consist of a web server, OpenDaylight controller and Software Defined Network emulated on Mininet as seen in Fig 2.

A. *Web Server*

The web server uses Ubuntu 17 as the operating system. We have configured Apache 2, which enabled us to host a simple html page as the website. This webserver is hosted on the Emulated Virtual Environment – Next Generation (EVE-NG) software. EVE-NG is the Emulated Virtual Environment for networking. It provides tools to be able to model a real-life network as virtual devices and interconnect them with other virtual or physical devices.

B. *OpenDaylight Controller*

In this project we used the eighth release of the OpenDaylight controller, which is called Oxygen. We downloaded the software from the OpenDaylight software download page and installed the controller on the Ubuntu 17 server. To install and enable required features that the OpenDaylight controller must use, an open source application called Apache Karaf is used. Karaf as it is normally called is a modular Open Services Gateway Initiative (OSGI) that provides tools and features required to deploy an application. An Open Services Gateway Initiative (OSGI) is a set of specifications for developing and deploying modular software programs and libraries, which are packed in bundles. Karaf enables modules to be installed, started, stopped, updated, and uninstalled without requiring a reboot.

By default, the OpenDaylight controller has no features enabled. We have installed and enabled the following features in this project on the controller (but there are many features, which can be installed and enabled)-:

- odl-restconf – Representational State (REST) like protocol that provides a programmatic interface over Hyper Text Transfer Protocol (HTTP) for accessing data on port 8080 for HTTP requests.
- odl-l2switch-all – Layer2 switch functionality.
- odl-mdsal-apidocs - Model Driven Service Abstraction Layer (MD-SAL) Application Programmable Interface (API) Documentation.
- odl-dlux-all - Graphical user interface for OpenDaylight based on the AngularJS Framework.

The OpenDaylight controller used on this paper is hosted on Amazon Web Services (AWS) cloud platform. We have used the Elastic Compute Cloud (EC2), which is a secure and resizable compute node. It allowed us to obtain and configure capacity in minutes.

The Elastic Compute Cloud (EC2) can scale both up and down allowing us to increase or decrease capacity as per our need. Wireshark, the network packet analyzer was used to analyze communication between the Open Virtual Switch in Mininet and the OpenDaylight controller.

When the communication between the switch and the controller is established. The controller adds flows to enable the switch to behave like a learning switch. When the switch receives a packet, it starts by performing a table lookup in the first flow table called table 0. In pipeline, each flow table contains one or more flow entries. Matching starts with the first flow table. If a Match is found, instructions associated with flow entry are executed. Instruction may direct the packet to next flow table in pipeline. When processing stops, the associated action set is applied, and packet forwarded. Instructions describe packet forwarding, packet modification, group table processing and pipeline processing. The summary flow chart is shown in Fig 3.

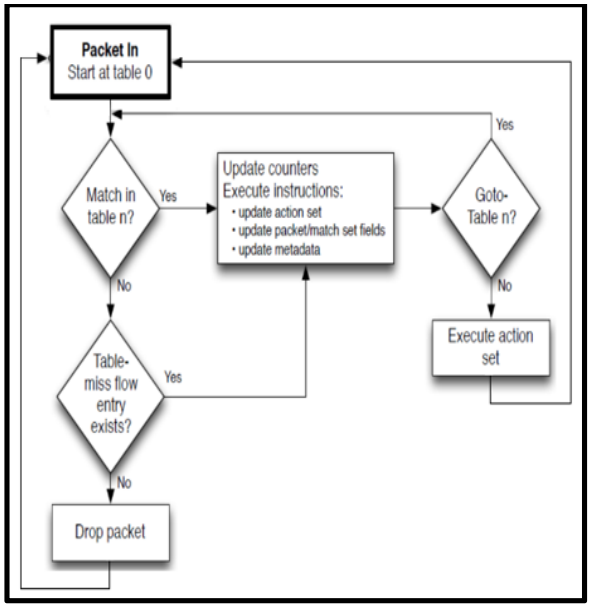


Fig. 3. Open Virtual Switch packet processing flow chart

C. Mininet

The Mininet Virtual Machine (VM) used in this paper is hosted on Oracles' Virtual Box. Mininet was installed on Ubuntu 17 operating system. The emulated Software Defined Network in Mininet was created using a Python script. The three main functions used in this script are Topo, Switch and Controller. Topo: This is the base class for Mininet topologies. It creates Data center network representation for structured multi-trees.

A function, which is used to create a custom network was created using Python. For this function to create the network, Mininet was prevented from creating the network using the default values. This was achieved by setting the topo class to none and the build class to false. Using the controller class, a remote controller was defined and given values for the name and an IP address, which in this case is the IP address of the OpenDaylight controller hosted on Amazon cloud. The connection port was set to port 6633.

Using OVSKernelSwitch sub class, an Open Virtual Switch (OVS) was created. Two (2) network hosts were also defined and given networking properties. The script defines the network subnet that the controller must use together with the links between the switch and the hosts. As part of the program, the script programs the controller to add the external interface to the switch after creating the network. This interface is used by Mininet to reach the Linux server inside the company domain.

To start the program, we loaded the saved python script from the directory that it was saved on. Mininet created the network as defined by the script. The created network consists of two (2) hosts and the Open Virtual Switch (OVS). The created switch has a control channel, which it uses to communicate with the controller. It has the pipeline, which consists of flow tables. There is also data path, which is the forwarding plane as seen in Fig 4.

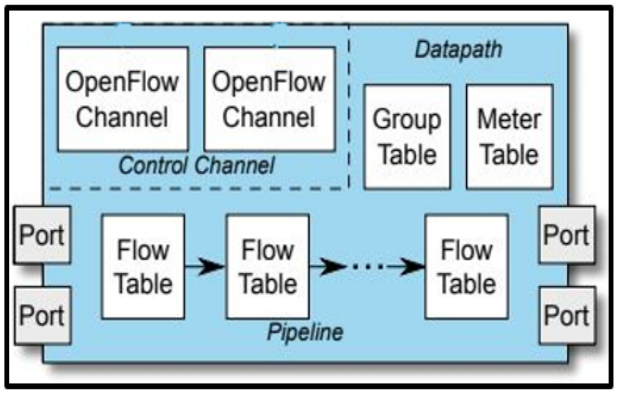


Fig. 4. Open Virtual Switch

D. Application Program Interface (API)

An application program interface (API) is a set of routines, protocols, and tools for building software applications. It specifies how software components should interact. They are used when programming graphical user interface (GUI) components. API makes it easier to develop a program by providing all the building blocks. One of the commonly used API in the field of networking is REST API.

Representational State Transfer (REST) API is an architectural style and an approach for communication used in the development of Web Services. It enables users to connect and interact with cloud services efficiently. To test the API, the program called Postman can be used. Postman is an application for testing APIs by sending request to the web server and getting the response back. Postman makes it easy to test, develop and document APIs. It allows users to set up all the headers and cookies the API expects and checks the response. "Postman" was used to send RESTCONF GET API to retrieve node inventory and topology as created by Mininet and seen by OpenDaylight controller.

RESTCONF is an Internet Engineering Task Force (IETF) draft that describes how to map a YANG specification to a RESTful interface. The REST-like API provide an additional simplified interface that follows REST-like principles and is compatible with a resource-oriented device abstraction. RESTCONF uses HTTP methods to provide CRUD (Create, read, update and delete) operations on a conceptual datastore containing YANG-defined data, which is compatible with a server that implements NETCONF datastores.

IV. RESULTS

"Iperf" is a tool used to test the maximum bandwidth that can be achieved between two (2) network devices. "Iperf" sends test data between the defined network devices and measures the throughput, bitrate, loss and other parameters. To test the functionality of the created network, an TCP "Iperf" test between the host and the Linux server was performed.

Using Wireshark, communication between the switch and the controller was captured. In the beginning of the communication, OpenFlow Channel messages between the switch and the controller are observed. The OpenDaylight requested the identity and basic capabilities of the switch. The switch responded with the requested information as seen with the OFPT_HELLO, OFPT_FEATURES_REQUEST and OFPT_FEATURES_REPLY packets as seen in Fig 5.

Source	Destination	Protocol	Lengt	Info
10.10.204.37	52.15.83.11	OpenFlow	74	Type: OFPT_HELLO
52.15.83.11	10.10.204.37	OpenFlow	90	Type: OFPT_FEATURES_REQUEST
10.10.204.37	52.15.83.11	OpenFlow	98	Type: OFPT_FEATURES_REPLY
52.15.83.11	10.10.204.37	OpenFlow	74	Type: OFPT_BARRIER_REQUEST
10.10.204.37	52.15.83.11	OpenFlow	74	Type: OFPT_BARRIER_REPLY
52.15.83.11	10.10.204.37	OpenFlow	82	Type: OFPT_MULTIPART_REQUEST, OFPMP_DESC
10.10.204.37	52.15.83.11	OpenFlow	1138	Type: OFPT_MULTIPART_REPLY, OFPMP_DESC
52.15.83.11	10.10.204.37	OpenFlow	114	Type: OFPT_MULTIPART_REQUEST, OFPMP_PORT_DESC
10.10.204.37	52.15.83.11	OpenFlow	98	Type: OFPT_MULTIPART_REPLY, OFPMP_METER_FEATURES
10.10.204.37	52.15.83.11	OpenFlow	122	Type: OFPT_MULTIPART_REPLY, OFPMP_GROUP_FEATURES
10.10.204.37	52.15.83.11	OpenFlow	338	Type: OFPT_MULTIPART_REPLY, OFPMP_PORT_DESC

Fig. 5. OpenFlow channel messages

Once the flows were added, the TCP "Iperf" test between host and the Linux server was successful. With the server using default TCP window size of 85.3 KBytes and the host using the default window size of 391 KByte, we were able to transfer 896 MBytes at a rate of 751 Mbits/sec from host to Linux server. From Linux server to host, 1.32 GBytes was transferred at a rate of 1.13 Gbits/sec as seen in the TCP "Iperf" results in Fig 6.

```
mininet-wifi> h1 iperf -c 10.1.10.2 -d
------------------------------------------------------------
Server listening on TCP port 5001
TCP window size: 85.3 KByte (default)
------------------------------------------------------------
------------------------------------------------------------
Client connecting to 10.1.10.2, TCP port 5001
TCP window size:  391 KByte (default)
------------------------------------------------------------
[  3] local 10.1.10.31 port 53638 connected with 10.1.10.2 port 5001
[  5] local 10.1.10.31 port 5001 connected with 10.1.10.2 port 51612
[ ID] Interval       Transfer     Bandwidth
[  3]  0.0-10.0 sec   896 MBytes   751 Mbits/sec
[  5]  0.0-10.0 sec  1.32 GBytes  1.13 Gbits/sec
mininet-wifi>
```

Fig. 6 Iperf test results

Another way of representing the Iperf test results using Wireshark is seen in Fig 7. Fig 7 shows the average throughput that will be the maximum bandwidth during the 90 ms period of the "Iperf" test for the uplink and downlink connections.

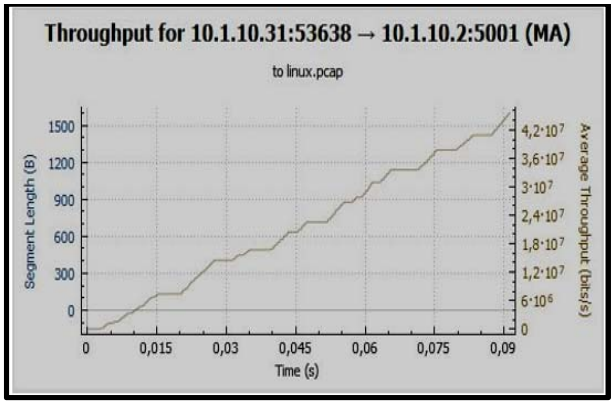


Fig. 7 Average throughput over 90ms

"Postman" was used to verify that indeed the created network in Fig. 8 is controlled by the OpenDaylight controller.

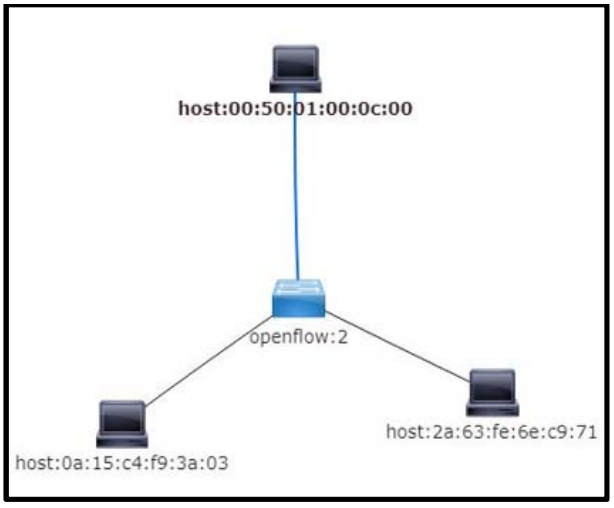


Fig. 8. ODL-DLUX network topology

To ensure reliability of the communication between the switch and the OpenDaylight controller, Transmission Control Protocol was used as seen in Fig. 9. The handshaking commands in the figure ensure reliability of the data transferred.

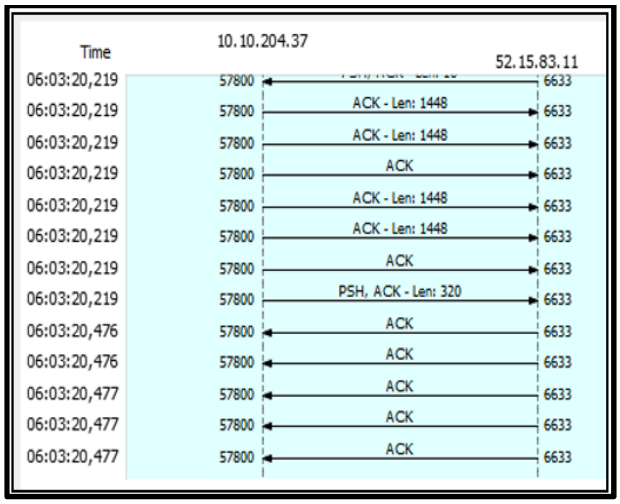


Fig. 9. TCP communication

V. CONCLUSION

In conclusion, Mininet was used to design a Software Defined Network. The SDN network was integrated to a real-life network using EVE-NG software via an OpenDaylight controller. OpenFlow protocol was modelled and used to facilitate communication between a virtual switch in Mininet and the OpenDaylight controller hosted in a cloud network. The OpenDaylight controller controls the flow of data from Mininet to the real-life network. The results showed the throughput and bandwidth measured in the communication process.

REFERENCES

[1] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, & J. Turner. "Openflow: enabling innovation in campus networks." ACM SIGCOMM Computer Communication Review, 38(2):69–74, 2008.

[2] Open Networking Foundation (ONF). Software defined networking: The new norm for networks. White Paper. From https://www.opennetworking.org/images/stories/downloads/sdn-resources/white-papers/wp-sdn-newnorm.pdf. 2012

[3] A. Lara, A. Kolasani, & B. Ramamurthy. "Network innovation using openflow: A survey". IEEE Communications Surveys and Tutorials, Vol 16 No 1 pp 1-20. 2013.

[4] H. Shimonishi, Y. Takamiya, Y. Chiba, K. Sugyo, Y. Hatano, K. Sonoda, K. Suzuki, D Kotani, & I. Akiyoshi. "Programmable network using OpenFlow for network researches and experiments". Proceedings of the Sixth International Conference on Mobile Computing and Ubiquitous Networking (pp.164-171). Okinawa, Japan. 2012

[5] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, & J. Turner, J. "OpenFlow: Enabling innovation in campus networks". ACM SIGCOMM Computer Communication Review, 38(2), 69–74. 2008

[6] A. Lara, A. Kolasani, & B. Ramamurthy. "Network innovation using OpenFlow: A survey". IEEE Communications Surveys and Tutorials, 16(1), 493–512. 2014

[7] C. C. Machado, L. Z. Granville, A. Schaeffer-Filho, & J. A. Wickboldt, "Towards SLA Policy Refinement for QoSManagement in Software-Defined Networking". IEEE 28th International Conference on Advanced Information Networking and Applications. 2014

[8] S. Sezer, S. Scott-Hayward, P. Chouhan, B. Fraser, D. Lake, J. Finnegan,N. Viljoen, M. Miller, & N. Rao. "Are we ready for sdn? Implementation challenges for software-defined networks". Communications Magazine, IEEE, vol. 51, no. 7, pp. 36–43, 2013.

[9] O. W. Paper. "Software-Defined Networking: The New Norm for Networks". Open Networking Foundation, Tech. Rep., April 2012.

[10] D. Thomas & N. K. Gray. "SDN: Software Defined Networks", Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472, First Edition, August 2013.

[11] J. Mambretti, J. Chen & F. Yeh. "Software-Defined Network Exchanges (SDXs): Architecture, Services, Capabilities, and Foundation Technologies". Proceedings of the 2014 26th International Teletraffic Congress (ITC). 2014

[12] B. Astuto, A. Nunes, M. Mendonca, X. Nguyen, K. Obraczka, & T Turletti. "A Survey of Software-Defined Networking: Past,Present, and Future of Programmable Networks". hal-00825087, version 5- 19 Jan. 2014.

[13] I. Z. Bholebawa & U. D. Dalal. "Design and Performance Analysis of OpenFlow-Enabled Network Topologies Using Mininet". International Journal of Computer and Communication Engineering. Volume 5, Number 6, November 2016.

[14] N. Handigol, B. Heller, V. Jeyakumar, B. Lantz & N. McKeown. "Reproducible Network Experiments Using Container-Based Emulation". CoNEXT'12, December 10–13, Nice, France. 2012

[15] Y. Huang, V. Jeyakumar, B. Lantz, B. O'Connor, N. Feamster, K. Winstein & A. Siyaraman. "Teaching Computer Networking with Mininet". Wi-Fi: HHonors, SGC. 2014

[16] J. Yan & D. Jin. "VT-Mininet: Virtual-time-enabled Mininet for Scalable and Accurate Software-Define Network Emulation". SOSR2015, Santa Clara, CA, USA. June 17–18, 2015.

[17] R. Kharga, P. Bholebawa, I. Satyarthi, S. Gupta, & S. Kumari. "OpenFlow technology: Ajourney of simulation tools". International Journal of Computer Network and Information Security, 6(11), 49–55. 2014

[18] F. Keti & S. Askar. "Emulation of Software Defined Networks Using Mininet in Different Simulation Environments". 6th International Conference on Intelligent Systems, Modelling and Simulation, Kuala Lumpur, 2015, pp. 205-210. 2016

[19] B. Lantz, B. Heller, & N. McKeown. "A network in a laptop: rapid prototyping for software-defined networks". in Proceedings of the 9thACM SIGCOMM Workshop on Hot Topics in Networks. ACM, 2010.

[20] K. K. Sharma & M. Sood. "Mininet as a Container Based Emulator for Software Defined Networks". International Journal of Advanced Research in Computer Science and Software Engineering. Volume 4, Issue 12, December 2014

[21] Python at https://www.python.org/

[22] M. Sanner. Python: "A Programming Language for Software Integration and Development". Article in Journal of Molecular Graphics and Modelling · November 1998

[23] OpenDaylight Consortium. http://www.opendaylight.org.

[24] W. Zhou, L. Li, M. Luo & W. Chou. "REST API Design Patterns for SDN Northbound API". 28th International Conference on Advanced Information Networking and Applications Workshops, Victoria, BC, 2014, pp. 358-365. 2014

[25] L. Li, W. Chou, W. Zhou & M. Luo. "Design Patterns and Extensibility of REST API for Networking Applications". IEEE Transactions on Network and Service Management, vol. 13, no. 1, pp. 154-167, March 2016.

[26] M. Brandt, R. Khondoker, R. Marx & K. Bayarou. "Security analysis of software defined networking protocols - OpenFlow, OF-Config and OVSDB". Paper presented at Fifth IEEE International Conference on Communications and Electronics, ICCE 2014, July 30 - August 1, 2014

[27] J. Schonwalder, M. Bjorklund & P. Shafer. "Network configuration management using NETCONF and YANG". IEEE Communications Magazine, vol. 48, no. 9, pp. 166-173, Sept. 2010

[28] Z. K. Khattak, M. Awais & A. Iqbal. "Performance evaluation of OpenDaylight SDN controller". 20th IEEE International Conference on Parallel and Distributed Systems (ICPADS), Hsinchu, 2014, pp. 671-676. 2014

[29] A. Mayoral, R. Vilalta, R. Muñoz, R. Casellas, R. Martínez & J. Vílchez. “Integrated IT and Network Orchestration using OpenStack OpenDaylight and Active Stateful PCE for Intra and Inter Data Center Connectivity”. ECOC 2014.

[30] S. Shin & G. Gu. CloudWatcher: “Network security monitoring using OpenFlow in dynamic cloud networks (or: How to provide security monitoring as a service in clouds?)”. 20th IEEE International Conference on Network Protocols (ICNP), Austin, TX, 2012, pp. 1-6. 2012

[31] E. Walker. “Benchmarking Amazon EC2 for High-Performance Scientic Computing”. USENIX login: Magazine, October 2008.

[32] I. Bermudez, S. Traverso, M. Mellia & M. Munafò, “Exploring the cloud from passive measurements: The Amazon AWS case”. Proceedings IEEE INFOCOM, Turin, 2013, pp. 230-234. 2013

Comprehensive Analysis of Routing Protocols for Cognitive Radio Ad-hoc Networks (CRAHNs)

Anand Nayyar
Graduate School
Duy Tan University
Da Nang, Vietnam
anandnayyar@duytan.edu.vn

***Abstract*—To resolve the issue of efficient spectrum allocation, a precious resource in communication systems, Cognitive Radio Adhoc Networks were proposed. Cognitive Radio Networks (CRNs) technology is becoming promising day by day as it combines technicalities of traditional ad hoc networks with Cognitive radio technology for efficient utilization of spectrum. Cognitive Radio Networks are surrounded by lots of technical challenges like spectrum variations in temporal and spatial domain, unknown behavior of primary and cognitive users, dynamic topology, security, routing and energy efficiency. Several protocols for optimizing routing and other parameters are proposed by researchers, but still lots more needs to be performed. The objective of this paper is to present detailed analysis of routing protocols in CRNs. The paper also presents a performance comparison of routing protocols to lay a strong base for researchers to find new scope for research to optimize CRNs.**

***Keywords*—Cognitive Radio Networks, Routing Protocols, Routing Challenges, RF Spectrum, Spectrum Opportunity (SOP), Routing Metrics**

I. INTRODUCTION

In recent years, radio frequency bands have demonstrated a rapid growth. In order to perform services, users make use of spectrum bands. The utilization of available frequency spectrum resources is also very inconsistent. There is a wide range of unlicensed bands which are overcrowded via utilization of growing technologies like Bluetooth, Infrared, Wi-Fi, WiMAX, etc. In comparison to unlicensed bands, licensed bands remain under-utilized [1]. Studies have shown that 5% of spectrum ranging from 30 MHz to 30GHz is only utilized in the United States. Considering the imbalance utilization of licensed and unlicensed Radio Frequency (RF) problem, an efficient solution is required to address the issue of spectrum allocation, where at certain frequencies, there exists lots of congestion and some bands are utilized inefficiently. To overcome this issue Cognitive radio technology is proposed.

Cognitive Radio technology is proposed with a strong vision of flexible utilization of RF spectrum. Cognitive radio technology will enable general users to operate in licensed frequency spectrums along with licensed or primary users. The mechanism by which secondary users can change their operating frequency is termed as Dynamic Spectrum Access (DSA). Cognitive radio networks have revolutionized wireless engineering with the usage of intelligent radios. With cognitive ratio network technology, radio devices can use radio frequencies in entirely new and efficient fashion. Cognitive radio networks have a special ability to monitor, sense and locate the operating conditions and reconfigure the characteristics to match with favorable conditions. Cognitive Radio technology is also termed as Hybrid Technology which constitutes Software Defined Radio (SDR). Cognitive radio provides transceiver with special abilities like: geographical location discovery, identification and user authorization, encrypting and decrypting signals, sensing nearby wireless devices and adjusting power and modulations.

Cognitive Radio is classified into two main categories: Full Cognitive Radio and Spectrum Sensing Cognitive Radio. In Full Cognitive Radio, all technical specifications of wireless devices are considered, whereas spectrum sensing cognitive radio is used to detect channels available in radio frequency spectrum.

CRN's comprise of wireless nodes with the CR capability with intelligent frequency spectrum sensing, radio reconfiguration and utilizing licensed spectrum bands which are available for usage. Such nodes are termed as "Secondary Users" / "Cognitive Users". In CRN Networks, the primary users are "Licensed Users" that can operate in licensed frequency bands. When any cognitive user wants to transmit any data, it can utilize the licensed frequency spectrum band which is not used by primary user. Therefore, the utmost responsibility of Cognitive user is to make use of licensed band in an *Opportunistic* manner without causing any sort of interference to primary users.

A Cognitive Adhoc Radio Network (CRAHN) [2] [3] has no centralized authority for controlling. Secondary users require special cooperation schemes to exchange crucial information like: Primary user's presence, configuration of node and free spectrum bands. In order to get this information, CRAHN requires efficient routing protocols. Currently, CRN based research is focused on PHY/MAC layer issues, but now research is also shifting towards multi-hop CRNs. CRAHNs are also known by unique characteristics like: Dynamic Topologies, Heterogeneity in spectrum, multi-hop architecture, self-configuration and energy efficiency. So, many researchers are doing a considerable amount of research in CRAHNs and are able to support diverse applications in operation with high efficiency.

Routing in CRAHNs is still highly unexplored area and requires immediate attention. Existing wireless networks, routing protocol cannot be directly utilized and integrated in CRNs, as these protocols can reduce overall efficiency and

can also make CRNs prone to failure. Some routing protocols proposed by researchers, especially for CRNs exist, but all the protocols are proposed on different objectives like: Maximization of spectrum opportunities, maximizing bandwidth availability, hop count minimization, end-to-end delay reduction, etc. Every routing protocol was tested on specific design targets and sometimes compared to just basic random routing protocols with different assumptions, settings and operational scenarios. But till date, no standard and dedicated routing protocol is proposed for CRNs with standard configuration. So, there is an ultimate requirement of the routing protocol to perform reliable and quick route discovery to boost overall CRN network.

The paper is organized as: Section II discusses Architecture of CRNs and Routing Metrics for Cognitive Radio Networks. Section III distinguishes Traditional Adhoc Networks with CRNs and lists various routing challenges surrounding CRNs. Section IV elaborates routing protocols for Cognitive Radio Networks (CRNs). Section V presents a tabular based performance comparison of routing protocols. Section VI concludes the paper with future scope.

II. Architecture & Radio Metrics for CRNs

Cognitive Radio Networks fall under into two main categories [4, 5, 9]:

1. With Infrastructure Cognitive Radio Networks
2. Without Infrastructure Cognitive Radio Networks

The components comprising the architecture of CRNs are enlisted as follows:

- With Infrastructure Cognitive Radio Network: In this type of network, the communication is via fixed infrastructure component termed as "CRN Base Station".
- Without Infrastructure Cognitive Radio Network: In this type of Network, the communication is done on the same lines as Adhoc multi-hop network.

Figure 1 gives a clear and detailed view of Architecture of CRNs.

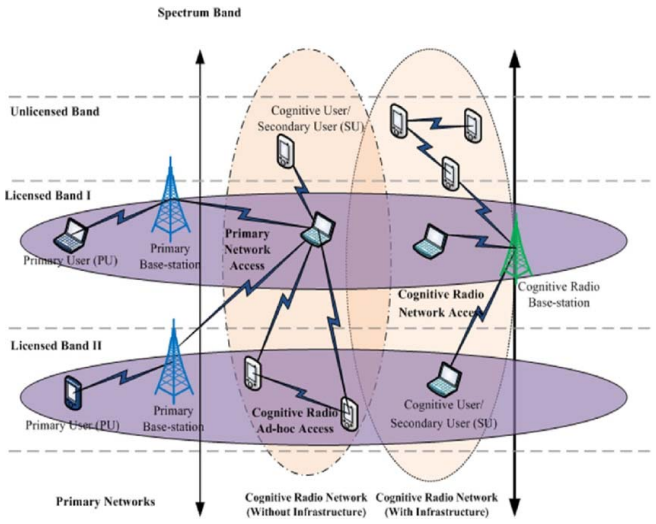


Fig. 1. CRNs: Architecture [10]

Primary Network: Primary network is regarded as a network which uses a separate RF spectrum frequency band for providing all sorts of services. Example: Cellular Tower, TV Broadcasting Network. Primary network consists of following components:

- ❖ Primary User: Primary user enjoys the main rights of the Licensed Spectrum. The primary user will not face any sort of interruption in communication via Secondary or Cognitive user.
- ❖ Primary Base Station: It is regarded as a backbone infrastructure component which controls all primary users. It doesn't have any capability to share any sort of frequency spectrum with secondary users.

Cognitive Network: These networks don't have any licensed RF spectrum band. The Cognitive Network consists of following components:

- Cognitive User: These users have a free band which doesn't come under licensed band. They can only make use of licensed bands in an opportunistic manner.
- Cognitive Base Station: It is regarded as a backbone infrastructure component for Cognitive users. And it provides a single hop connection to secondary users.

Routing Metrics for Cognitive Radio Networks

For efficient transmission of data from source to destination, every routing protocol uses one or more metrics to lay down the efficient path. Routing metrics are defined as the parameters utilized by routing protocols to perform routing in the network. Routing metrics lay a strong foundation for protocol to follow or reject the path while doing data transfer. So, routing metrics are baselines to determine the quality control of the network.

The following are the Routing Metrics in CRNs are classified as follows:

- End-to-End Delay: It is calculated on the basis of the transmission time of packet transfer between source and destination node. End-to-End delay depends on varied delay parameters like: Propagation, Switching and Queuing.
- Energy Efficiency: Energy consumed by wireless nodes for transmitting packets from source to destination, i.e. transmitting the packet from node to node, receipt of packet and switching of packet.
- Bandwidth: Bandwidth is data transfer rate at which packets are transmitted from source to destination. Higher bandwidth means better routing path.
- Link Stability: It refers to the stability of the route required to transfer the packet and also channel availability.

- Hop Count: Hop count determines the number of nodes, the packet has to traverse from source to destination node. It doesn't depend on channel quality. Lower hop count means less nodes in path and packet will reach the destination quickly.

III. Cognitive Adhoc Radio Networks V/S Traditional Adhoc Networks

The following points highlight the differences between Cognitive Radio Networks and Traditional Adhoc Networks:

- Selection of Transmission Spectrum: In traditional Adhoc networks [21], the users transmit the data over pre-selected channel and it doesn't get changed over time. Traditional Adhoc networks have support for multiple-channels which are available for users to transmit and nodes can select any channel in a random manner. Considering CRNs, the spectrum bands are available and vary with time and space. Primary users can select the Licensed whereas cognitive users enjoy unlicensed but cognitive users can make use of Licensed spectrum in an opportunistic manner.

- Topology: Adhoc networks have no centralized coordinating entity and nodes rely on local coordination to formulate topology. In comparison to CRNs, licensed spectrum opportunity exists and beacons are sent out to explore all possible channels. CRNs may face a challenge, i.e. lack of information in terms of topology which leads to collisions among primary users and cognitive users.

- Multi-Hop Transmission: In traditional Adhoc networks getting proper End-to-End QoS requires lots of traffic load, varied channels and frequency bands. In CRNs, end-to-end QoS is possible via routing and spectrum allocation collaboration.

CRNs Routing Challenges

The following points enlists various routing challenges surrounding CRNs:

- Networking Issues: Routing protocols proposed for wired and wireless networks are not at all suitable for Cognitive Radio networks due to dynamic topology changes. These dynamic topology changes bring new challenges in network operations in terms of mobility, limited power, network lifetime and attenuation with other ISM-devices.

- Spectrum Operational Issues: Primary task of Cognitive radio networks is to make sure for the availability of spectrum for communication. Routing metric has to carefully analyze the spectrum range on the basis of availability so that transmission should not affect between primary or secondary users. In real time operational environments, Cognitive Radio Networks have lots of contrasting channel properties like bandwidth, delay, loss rate, propagation characteristics etc.

- Energy Utilization: In CRNs, secondary users make use of less licensed spectrum and have energy utilization restrictions. Secondary users should be capable to sense the spectrum in the high frequency band, so efficient power control schemes are required.

- Deafness Issue: Deafness problem is one of the serious issues surrounding routing in CRNs. Considering multi-channel communication, when a sensor node is making use of one channel band for sending and receiving data cannot sense another channel signaling. The most obvious solution to this deafness problem is to make use of Common Control Channel (CCC) shared among all operational nodes for performing the initialization of routes and data maintenance. But, this is not an efficient solution. Other solution can be utilization of channel synchronization schemes, but this can lead to delay issues and high-power consumption. Deafness problem affects routing selection and even overall performance of the network.

- Security Issue: Another serious issue surrounding CRNs is security. In CRNs, secondary users make use of licensed spectrum, which is only made for primary users via spectrum sensing technology. Because of this access, attackers can even have access to licensed spectrum and attackers can launch various sorts of attacks like DoS Attack, Emulation attack causing serious impact on CRN operations. In addition to this, CRNs can even face other network threats like Forgery, Tampering, Eavesdropping, Malicious node infections, etc. So, security has to be optimized for efficient routing.

IV. CRNs Routing Protocols

The following are the routing protocols for CRNs [6,7,8,11]

A. Dynamic Spectrum & On Demand -Aware Routing

Dynamic Spectrum Aware routing protocols are designed with an objective to make efficient utilization of unallocated wireless spectrum. In these routing protocols, discovery of route take place on the basis of spectrum sensing. These protocols determine and establish efficient routes across various other routes in available spectrum.

Dynamic Spectrum Aware Routing Protocols are:

- SPEAR
- SAMER
- SORP
- MSCRP

SPEctrum-Aware Routing (SPEAR) [12]: Sampath et al. proposed SPEctrum Aware Routing Protocol. SPEAR protocol provides efficient packet transmission in terms of throughput during spectrum heterogeneity presence. It

integrates flow-based approaches and link-based approaches for providing end-to-end connectivity. SPEAR protocol has unique characteristics in terms of: Integration of spectrum discovery with route discovery to cope with spectrum heterogeneity to efficiently utilize channels; Coordination of channel utilization across nodes to optimize the channel assignment on the basis of per-flow and minimize inter-flow interference; Assigning different channels to link on the same flow to reduce intra-flow interference. In SPEAR routing protocol, discovery of the route is done via Route Request (RREQ) message broadcast on a dedicated control channel and is identified by source and destination IP addresses. Unlike traditional routing protocols like AODV, SPEAR discovers multiple paths for the destination. The destination node makes optimal path selection on the basis of throughput and End-to-End delay.

SPEAR protocol was tested on Qualnet Simulator on the 1000m x 1000m grid, 12Mbps data rate of Cognitive and Control Radios using 802.11 CSMA/CA MAC protocol. Results state that SPEAR protocol outperforms link and flow-based approaches in terms of throughput and hop count.

Spectrum Aware MEsh Routing (SAMER) [13]: Pefkianakis et al. proposed SAMER routing protocol. The objective of SAMER protocol is to make efficient use of network spectrum in an opportunistic manner by routing traffic among different paths with high spectrum availability and long-term stability. SAMER protocol uses two-tier routing approach to maintain proper balance between hop count and spectrum availability. The two components that make up SAMER Protocol are:

1. Dynamic Candidate Mesh- To build the path between transmitting nodes with shortest hop count and adaptable to all changes in the spectrum.

2. Opportunistic Forwarding: Forwarding the traffic in opportunistic manner to those links having high spectrum.

The SAMER protocol was tested using Qualnet simulator on 1500m x 1500m terrain, on 10Mbps frequency band, divided by 5 Spectrum blocks of 2 Mbps. Simulation results proves that SAMER protocol has high throughput and performs efficiently in CORNET.

Spectrum-aware On-Demand Routing Protocol (SORP) [14]: Cheng et al. proposed SORP routing protocol for efficient route discovery and spectrum band selection. The protocol inherits basic routing functionality from AODV protocol with modified Route Request (RREQ). In SORP protocol, all the SOP information is piggybacked by RREQ messages and is only forwarded when the node finds any sort of interaction between RREQ's and its own. The destination node knows SOP distribution when RREQ is received. It transmits back a RREP message to the source and intermediate nodes gets frequency band via band choices from RREP.

SORP routing protocol was tested using GloMoSim simulator on 1800 m x 1800 m area, radio range of 372.214 m, 100 wireless nodes with every node having two interfaces. The results state that SORP outperforms AODV and other traditional networks and gives less cumulative delay.

Multi-Hop Single-Transceiver Cognitive Radio Networks Routing Protocol (MSCRP) [15]: Ma et al. proposed MSCRP routing protocol. In this protocol, AODV protocol is modified for exchanging the channel information among all nodes operational in the network. Route discovery is done via broadcasting RREQ message, piggybacking channel information. The protocol doesn't make use of the Common Control Channel. In case of data transmission, the node broadcasts a RREQ packet as nodes are on different channels. When any intermediate node receives RREQ, it transmits the packets also available on all channels. The channel information is piggybacked by RREQ messages to broadcast process. MSCRP protocol uses LEAVE/JOIN message to inform neighboring nodes regarding channel.

MSCRP protocol was tested on OPNET simulator with 12 and 6 licensed channels, AODV routing protocol, 24 Nodes, flow rate of 100 Kbps to 1700 Kbps. Results state that MSCRP outshines AODV routing protocol in terms of throughput.

B. Tree Based Routing Protocols

In order to optimize Cognitive Radio Networks, tree-based routing protocols come into the picture. In tree-based routing protocols, centralized routing scheme is followed, which is controlled via single networking entity called "Base Station". Tree BASED Routing Protocols are: CTBR and STOD-RP.

Cognitive Tree-based Routing (CTBR) [16]: Zhang et al. proposed CTBR protocol as an advanced extension towards TBR protocol to support multiple wireless systems like IEEE 802.11g and 802.11j. Traditional Tree Based Routing protocol was efficient protocol and implemented mainly when network is configured with base station, but cannot be applied directly to CRNs as it cannot handle multiple wireless systems. In order to make TBR work for CRN's it is added with novel cognitive-aware link metric for enhancing link quality. A global decision scheme is implemented for route selection with efficient end-to-end metric. In CTBR, root broadcasts Root Announcement (RANN) message at regular intervals to create a tree. When a Cognitive Transceiver (CT) receives the message, it caches the CT from whom it receives the announcement message as its parent. After that, CT does re-broadcasting of RANN message with updated link metric. In addition to this, CT selects the best parent on the basis of best metric from the source to root. In order to be registered with Root, every CT broadcasts RREP message towards the root node on receiving RANN message. The routing table is also updated by selecting source CT of RREP at its destination node. Finally, the tree is constructed in which every node can be reached on the network.

CTBR routing protocol was tested on OPNET simulator on 500m x 500m area, RANN broadcast interval of 5seconds, 54 Mbps Transmission bit rate, 120m & 60 m transmission range using Two-ray propagation model. Simulation results prove that CTBR has less End-to-End delay and high packet delivery ratio, less routing overhead and performs well in CRN environments.

Spectrum-Tree based On-Demand Routing Protocol (STOD-RP) [17]: Zhu et al. proposed STOD-RP routing protocol for CRN via designing a spectrum tree in every spectrum band using the new route metric. In STOD-RP protocol, every node is fitted with spectrum agile radio for doing tasks like spectrum awareness and reconfiguration. All the nodes perform packet forwarding in ad hoc manner. In this protocol, a new route metric is proposed by determining the activities of primary user and QoS requirements of secondary user. On the basis of that, the spectrum-tree is constructed and end-to-end route is determined using STOD-RP protocol. STOD-RA uses a mixed approach of routing via Tree-based proactive routing and on-demand route discovery. It is an extended version of AODV routing protocol and also classifies routing on basis of intra-spectrum and inter-spectrum. It uses SSREQ and SRREP messages to discover the routing paths between source and destination nodes. SSREQ messages is extended version of standard RREQ and contains CRIDS, CRIDD, Metric, intra/inter. SSREP message contains CRIDS, CRIDD and intra/inter.

STOD-RP protocol was tested on NS-2 simulator on 500 m x 500m area, bit rate at 2 Mbps, transmission range of 250m and simulation time of 150 seconds. The protocol is compared with CTBR protocol and the results states that STOD-RP protocol outshines in end-to-end delay performance and has less routing overhead.

B. Local Coordination based Routing

Another efficient routing approach in Cognitive Radio Networks is Local Coordination based. It provides efficient solutions for handling inconsistencies in spectrum bands.

Local Coordination based Routing and Spectrum Assignment in Multi-Hop Cognitive Radio Networks [18]: The protocol is primarily an extension to SORP routing protocol involving integration of a local coordination scheme to apply towards intersecting nodes to perform the task of load balancing. In this approach, AODV routing protocol is improvised to create a novel mechanism on CCC for exchange of spectrum opportunity to overcome the problem of inconsistency. The protocol consists of two parts: Routing Assignment and Spectrum Assignment. The protocol establishes multi-hop path and assigns appropriate frequency band to each hop. The protocol exchanges SOP information between network nodes and assign best spectrum band to every link on the route established. The local coordination scheme gets invoked on the node it converts the node into intersecting node. The protocol consists of following modules:

- Multi-Flow multi-frequency scheduling: Under this, nodes perform polling among the active frequency bands and use specific frequencies at least one-time during polling cycle processing.
- On-demand routing: Under this appropriate frequency band is selected.
- Local coordination Scheme: The scheme enables the node to decide whether to perform flow accommodation or flow redirection.

The protocol was tested on MATLAB workbench using 100 wireless nodes at 1800 m x 1800 m area, radio range of 372.214 m and 6 accessible frequency bands. Simulation results state that the protocol is highly efficient in spectrum allocation, less end-to-end delay and outperforms all traditional routing protocols in terms of routing.

C. Multipath Routing Protocols

Under this category of routing protocols, multiple paths from source to destination nodes are discovered and best paths are selected for transmission. As compared to other routing protocols, multi-path routing protocols have add-on advantages in terms of end-to-end delay, better throughput, energy efficiency and overhead. The following protocol comes under Multi-path routing protocol category for CRN: Multipath Routing and Spectrum Access (MRSA).

Multipath Routing and Spectrum Access (MRSA) routing protocol [19]: Wang et al. proposed MRSA, for reducing contention and interference among stations with regard to load balancing. The protocol overcomes all interruptions of primary users by traffic distribution over multiple paths. MRSA protocol makes assumption of N bands for sending packets across the spectrum.

MRSA routing is based on DSR technique, in which source node broadcasts RREQ message with new RREQ_ID along with its own BRT. All intermediate nodes on getting RREQ messages checks the RREQ message, whether the ID in the message is new or old. If the ID is old, the hop count is calculated by the relay node from the source node. If the ID is new, the node attaches BRT to incoming RREQ and RREQ is rebroadcasted in the network.

MRSA protocol was tested using NS2 simulator [20, 22] on 100 m x 100m, 20 secondary users, 5 primary Users and transmission range of 25 meters. MRSA protocol was compared with SMR and SPMR and the results prove that MRSA protocol outperforms in End-to-End delay, throughput and has better link stability as compared to SMR and SPMR protocols.

V. Performance Comparison of Routing Protocols for CRNs

Table 1 enlists the performance comparison of Routing Protocols in Cognitive Radio Networks (CRN) on the basis of: Type of Protocol, Discovery of Route, Routing Decision, Selection of Best Path, Network Topology, Simulator Testbed, Parameters Optimized for CRN.

VI. Conclusion & Future Scope

Cognitive radio networks (CRNs) are surrounded by lots of challenges and routing is regarded as one the serious challenge which is required to be solved due to diversifications in channel availability, data transmission rates. Researchers across the globe are working on finding some novel solutions, especially with regard to routing to make CRNs more reliable and efficient network in spectrum distribution between Primary Users and Cognitive Users. This paper discuss architecture, routing metrics, routing challenges faced by Cognitive Radio Networks and highlights common differences between CRN's and Traditional Adhoc Networks are also listed. The paper presents a comprehensive survey of various routing protocols of CRNs based on Dynamic Spectrum, Tree based, Local Coordination and Multipath. In addition to this, the performance comparison of routing protocols on the basis of discovery of route, routing decision, best path selection, type of protocol, network topology, simulator utilized for testing and parameters optimized by routing protocol for improvising cognitive radio networks is also enlightened.

In the near future, we like to explore more details regarding the technical specifications of routing protocols and figure out some gaps in operations to propose a new routing metric cum routing protocol highly efficient towards: spectrum allocation, security and high data rate transmission with low latency and QoS with regard to end-to-end delay. The protocol will be tested using NS-2 or NetSim simulator for performance evaluation.

References

[1] Fette, B. A. (Ed.). (2006). *Cognitive radio technology*. Newnes.

[2] Zhang, Y., Zheng, J., & Chen, H. H. (Eds.). (2016). *Cognitive radio networks: architectures, protocols, and standards*. CRC press.

[3] Ahmed, Z. I. A. M., Bilal, D. K. H., & Alhassan, D. M. (2016). Cognitive Radio Network Review. *International Journal of Engineering", Applied and Management Sciences Paradigms*, *13*(1), 320-327.

[4] Ali, A., Iqbal, M., Baig, A., & Wang, X. (2011). Routing techniques in cognitive radio networks: a survey. *International Journal of Wireless & Mobile Networks*, *3*(3), 96-110.

[5] Banerjee, J. S., Chakraborty, A., & Karmakar, K. (2013). Architecture of Cognitive Radio Networks. *Cognitive Radio Technology Applications for Wireless and Mobile Ad Hoc Networks*, 125-152.

[6] Cesana, M., Cuomo, F., & Ekici, E. (2011). Routing in cognitive radio networks: Challenges and solutions. *Ad Hoc Networks*, *9*(3), 228-248.

[7] Abdelaziza, S., & El Nainay, M. (2012). Survey of routing protocols in cognitive radio networks. *Preprint submitted to Elsevier*, 1-20.

[8] Akyildiz, I. F., Lee, W. Y., & Chowdhury, K. R. (2009). CRAHNs: Cognitive radio ad hoc networks. *AD hoc networks*, *7*(5), 810-836.

[9] Mansoor, N., Islam, A. M., Zareei, M., Baharun, S., Wakabayashi, T., & Komaki, S. (2015). Cognitive radio ad-hoc network architectures: a survey. *Wireless Personal Communications*, *81*(3), 1117-1142.

[10] Akyildiz, I. F., Lee, W. Y., Vuran, M. C., & Mohanty, S. (2006). NeXt generation/dynamic spectrum access/cognitive radio wireless networks: A survey. *Computer networks*, *50*(13), 2127-2159.

[11] Amjad, M., Rehmani, M. H., & Mao, S. (2018). Wireless multimedia cognitive radio networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, *20*(2), 1056-1103.

[12] Sampath, A., Yang, L., Cao, L., Zheng, H., & Zhao, B. Y. (2008). High throughput spectrum-aware routing for cognitive radio networks. *Proc. of IEEE Crowncom*.

[13] Pefkianakis, I., Wong, S. H., & Lu, S. (2008, October). SAMER: Spectrum aware mesh routing in cognitive radio networks. In *New Frontiers in Dynamic Spectrum Access Networks, 2008. DySPAN 2008. 3rd IEEE Symposium on* (pp. 1-5). IEEE.

[14] Cheng, G., Liu, W., Li, Y., & Cheng, W. (2007, April). Spectrum aware on-demand routing in cognitive radio networks. In *New Frontiers in Dynamic Spectrum Access Networks, 2007. DySPAN 2007. 2nd IEEE International Symposium on* (pp. 571-574). IEEE.

[15] Ma, H., Zheng, L., & Ma, X. (2008, May). Spectrum aware routing for multi-hop cognitive radio networks with a single transceiver. In *Cognitive Radio Oriented Wireless Networks and Communications, 2008. CrownCom 2008. 3rd International Conference on* (pp. 1-6). IEEE.

[16] Zhang, B., Takizawa, Y., Hasagawa, A., Yamaguchi, A., & Obana, S. (2007, March). Tree-based routing protocol for cognitive wireless access networks. In *Wireless Communications and Networking Conference, 2007. WCNC 2007. IEEE* (pp. 4204-4208). IEEE.

[17] Zhu, G. M., Akyildiz, I. F., & Kuo, G. S. (2008, November). STOD-RP: A spectrum-tree based on-demand routing protocol for multi-hop cognitive radio networks. In *Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008. IEEE* (pp. 1-5). IEEE.

[18] Yang, Z., Cheng, G., Liu, W., Yuan, W., & Cheng, W. (2008). Local coordination-based routing and spectrum assignment in multi-hop cognitive radio networks. *Mobile Networks and Applications*, *13*(1-2), 67-81.

[19] Wang, X., Kwon, T. T., & Choi, Y. (2009, September). A multipath routing and spectrum access (MRSA) framework for cognitive radio systems in multi-radio mesh networks. In *Proceedings of the 2009 ACM workshop on Cognitive radio networks* (pp. 55-60). ACM.

[20] Nayyar, A., & Singh, R. (2015). A comprehensive review of simulation tools for wireless sensor networks (WSNs). *Journal of Wireless Networking and Communications*, *5*(1), 19-47.

[21] Kaur, M., & Nayyar, A. (2013). A comprehensive review of mobile adhoc networks (MANETS). *International journal of emerging trends & technology in computer science (IJETTCS)*, *2*(6).

[22] Nayyar, A., & Singh, R. (2014). A comprehensive review of ant colony optimization (ACO) based energy-efficient routing protocols for wireless sensor networks. *International Journal of Wireless Networks and Broadband Technologies (IJWNBT)*, *3*(3), 33-55.

TABLE 1: PERFORMANCE COMPARISON OF ROUTING PROTOCOLS OF COGNITIVE RADIO AD-HOC NETWORKS (CRAHNS)

Name of Protocol	Discovery of Route	Routing Decision	Selection of Best Path	Protocol Type	Network Topology	Simulator Testbed	Parameters Optimized
SPEAR	Common Control Channel, Broadcast of RREQ Message	Robust Paths	Integration of Flow and Link based approaches	On-Demand	Linear Topology	Qualnet	Throughput
SAMER	Link State Packets	Collaboration of PHY and MAC layers	Minimum Hop Count and Availability of Spectrum	Dynamic Spectrum Aware	Mesh Topology	Qualnet	Efficient distribution of data traffic across paths with under-utilized spectrum, End-to-End throughput
SORP	RREQ Message broadcast	Collaboration of MAC and Network Layers	Switching Delay and Back Off Delay	On Demand	Mesh Topology	GloMoSim	Cumulative Delay
MSCRP	RREQ Message Broadcast	Collaboration of MAC and Network Layers	Number of flows on each channel	On Demand	Star Topology	OPNET	Throughput
CTBR	Broadcast Root Announcement (RANN)	Don't based on Cross Layer	Updation of single field by every node "Cumulative Metric"	Tree Based	Tree Topology	OPNET	Packet Delivery Ratio, End-to-End Delay, Routing Overhead, Average Hop Count
STOD-RP	SSREQ with Updated field like CRIDS, CRIDD, Metric, intra/inter	Construction of Spectrum Tree in every spectrum band by considering "New Route Metric"	Cognitive Route Metric Computation	Tree Based	Tree Topology	NS-2	End-to-End Delay, Routing Overhead
Local Coordination	Broadcasting RREQ message	Joint decisions based on MAC and Network Layer	Based on Cumulative delay of the path	Local Coordination	Mesh Topology	MATLAB	End-to-End Delay
MRSA	Broadcasting of RREQ Message with new ID	Not based on Cross Layer	Minimum Hop Count	Multipath Routing	Simple and Random Topology	NS2	End-to-End delay, Throughput, Link Stability

Analysis on Modulation Techniques of an AC drive with Respect to Harmonic Content and Efficiency

Josias W. Makhubele
Department of Electrical & Electronics Engineering Technology
Faculty of Engineering and the Built Environment
University of Johannesburg
Johannesburg, South Africa
201038210@student.uj.ac.za

Kingsley A. Ogudo
Department of Electrical & Electronics Engineering Technology
Faculty of Engineering and the Built Environment
University of Johannesburg
Johannesburg, South Africa
kingsleyo@uj.ac.za

***Abstract*—Inverters have become popularly used in the field of Power Electronics for different applications. Inverters having harmonics existence in the output voltage have led to many research works being conducted. Different topologies were applied to mitigate this inverter drawback, thus improving the efficiency of the inverter. Sinusoidal Pulse Width Modulation (SPWM), Hysteresis Band Current Controller Pulse Width Modulation (HCCPWM) and Modified Sinusoidal Pulse Width Modulation (MSPWM) techniques have been discussed and implemented in Matlab/Simulink software. Full-bridge single phase topology is used. Simulations results are presented. Modulation index and number of pulses are only two usual parameters used in these three modulation techniques. Minimum Total Harmonic Distortion (THD) is applied as a benchmark key performance indicator (KPI) in comparing the output results generated from all the techniques investigated and verified. The results show that Modified Sinusoidal PWM technique produce better results as compared to SPWM and HCCPWM techniques. The results generated from MSPWM technique have reduced harmonics and the output voltage is more sinusoidal without the use of an ideal switching technique. This paper demonstrated the output voltage performance relation between SPWM, HCC and MSPWM carried-out in the unipolar inverter.**

Keywords—Sinusoidal Pulse Width (SPWM), Hysteresis-band Current Control (HCC), Space Vector Pulse Width (SVPWM), Selective Harmonic Elimination (SHE),Phase Shifted Carrier PWM (PSCPWM) and Pulse Width Modulation (PWM)

I. INTRODUCTION

In the modern years, multiphase voltage source inverter (VSI) have been generally utilized for multi-phase induction motor for variable frequency and variable voltage. Applications such as locomotive traction, electric aircraft, electric ship propulsion, electric and hybrid electric vehicles and other high-power electronics. VSI is a step-down type of an inverter and use PWM techniques. Current Source Inverter (CSI) is not mostly used due to commutation complexity in the circuit. Inventive approaches have been advanced to overcome the poor application of input DC voltage. The arrangement of Pulse Width Modulation (PWM) can increase a fundamental component in an intelligent manner as it has been discovered. An increased fundamental output voltage and reduced harmonic contents results from an enlarged width of the pulse at the center of the wave [5, 7 & 8]. In the past years, research work has been conducted by different researchers on the reduction of harmonic content in the output voltage and current waveform of the inverters [6]. Reduced load efficiency and performance may result from harmonics contained in the inverter output waveforms (either voltage or current) that are rectangular in nature [2]. A.C. drives are called Inverters and are mostly used for industrial applications for speed control of a motor mechanism. Different modulation techniques are used to eliminate harmonics in A.C. drives. These modulation techniques have different algorithms. Any modulation technique serves a purpose of producing variable output with a maximum fundamental component having minimum harmonics [9 &10]. To control the switching pattern of the inverter different PWM techniques are applied such as; Sinusoidal Pulse Width (SPWM), Space Vector PWM (SVPWM), Selective Harmonic Elimination (SHE), Hysteresis Band Current Control (HCC PWM), Trapezoidal Pulse Width Modulation Technique (TPWM), Third Harmonic Injection Pulse Width Modulation (THIPWM). This paper will only focus on SPWM, HCC and Modified SPWM and results will be analyzed.

II. MODULATION TECHNIQUES

A. Sinusoidal PWM technique

Sinusoidal Pulse Width Modulation Technique is mostly applied in Power Electronics. Power converters uses mostly SPWM technique due to its advantage of easy to produce good output waveform with harmonic content that are tightly controlled. The main drawback of SPWM is harmonic amplitudes of the energy associated with the switching frequency and multipliers are very high [1, 3, 8, 10, 13, 14 & 15]. Carrier-based SPWM has a very low complexity and good dynamic response for matrix converters as its main advantage [16]. SPWM has two classifications such as: Bipolar and Unipolar techniques. Gating signals are generated based on the comparison between sine wave and triangular wave in Bipolar SPWM technique. Second technique is Unipolar SPWM, this method has two sinusoidal waveforms with given offset that are matched-up with a triangular waveform and based on matching-up the gating signals are produced for internal control of the

inverter [1, 4, 10 & 12]. The SPWM can provide gate triggering signal for controlling the IGBT or MOSFET switches in power converters. Modulation index "m" is the fundamental parameter that is needed to design SPWM. Amplitude of the output voltage can be varied by varying this parameter. Figure 1 show the creation of gate pulses by SPWM.

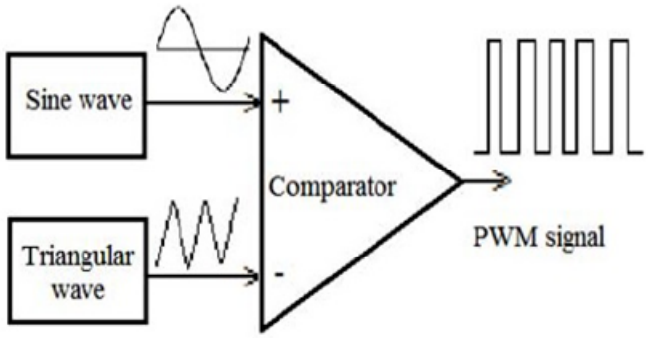


Figure 1: PWM signal generating [24].

Reference signal frequency determines the output frequency of the inverter. The reference signal frequency can be applied to the carrier frequency as well as to identify the number of pulses per fractional cycle.

$$n_p = \frac{f_c}{f_r} \qquad (1)$$

Where, n_p is the number of pulses per fractional cycle. Modulation index as defined below is given by the proportionality between the amplitude of the reference voltage V_r to the amplitude of the carrier voltage signal V_c :

$$m = \frac{V_r}{V_c} \qquad (2)$$

PWM invented with desired m ,output voltage waveform can be generated with reduced THD [23]. SPWM gets generated when two reference signals (one being 180 degrees phase shifted) and having equal size and frequency are compared with large frequency signal in this case is operated by Unipolar inverter switching scheme [24]. Inverters that use PWM switching techniques have DC input voltage that is usually constant in size [25].

B. Hysteresis –Band Current Control Pulse Width Modulation (HCCPWM)

Hysteresis-band Current Controller is a popular PWM technique because of its application integrity, fast transient response, direct limiting of device peak current. HCC PWM control method usually requires voltage and current sensors for successful implementation in almost all conditions including unbalanced conditions. HCC is essentially a spontaneous reaction current control technique of PWM where definite current tracks the command current continuously within the hysteresis band [11&19-21]. The aim of HCC is to keep outside current of the converter within hysteresis band determined. The advantages of this technique are simple control algorithm, Dynamic response, good accuracy, robustness and robustness, simple in design and need for information about load parameters required is not necessary [15, 17 & 18].

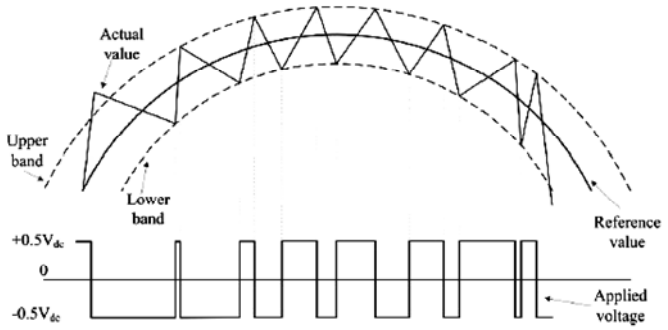


Figure 2: Basis of Hysteresis Band Current Control [26].

The application procedure of Hysteresis Band Current Control for a half bridge inverter is shown in Figure 2. The control circuit generates the sine reference current wave. The output voltage changeover from $+0.5V_d$ to $-0.5V_d$ is obtained from the half-bridge, when upper switch is turned off and lower switch is turned on, the current outruns the prescribed hysteresis band and current then starts to decay. When upper switch is turned on, the current crosses the lower band limit and then turns off the lower switch. Backwards and forwards switching of upper and lower switches forces the actual current wave to follow the Sine reference wave within the hysteresis band. Irrespective of V_d fluctuation, an inverter basically becomes a current source with peak-to-peak ripple current which is regulated within a hysteresis band. Hysteresis band width is dependent to ripple current peak-to-peak and switching frequency. The following algorithm applied to hysteresis current controller gives out pulses to the VSI inverter: $|i_{m,ref} - i_m| < \varepsilon$ keeps the output pulse at the same state $|i_{m,ref} - i_m| > \varepsilon$ let output pulse = 1 (High) $|i_{m,ref} - i_m| < -\varepsilon$ let output pulse = 0 (Low) and where m = a,b,c phases and ε is the hysteresis band [26].

The equations or derivations of this scheme are:

$i_{ref(t)} = i_{m,ref} \sin(\omega t)$

Upper band $i_u = i_{m,ref(t)} + \Delta i$ (1)

Lower band $i_l = i_{m,ref(t)} - \Delta i$ (2)

Where Δi =hysteresis band limit

If $im > iu$, $vm = \frac{\text{vdc}}{2}$ (3)

If $im < iu$, $vm = \frac{\text{-vdc}}{2}$ (4)

i is load current and V_{dc} is the dc link voltage of the inverter.

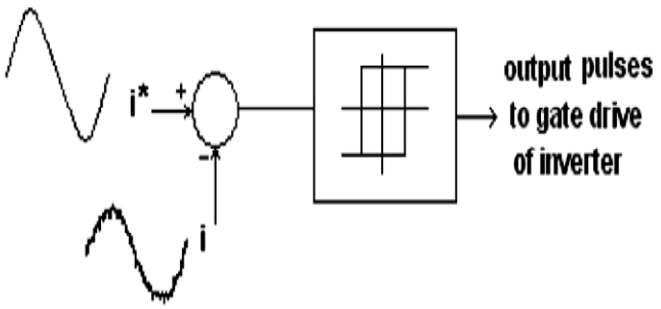


Figure 3: HBPWM Control block diagram [26].

Hysteresis comparators are utilised to force a hysteresis band from the reference current. Therefore, excellent dynamic performance is provided by this control scheme because of its quick action capability. None-optimum harmonics and unstable PWM frequency are the main drawbacks of this technique. [26&27].

C. Modified Sinusoidal Pulse Width Modulation(MSPWM)

Modified SPWM technique is very similar to traditional SPWM, where gating signals are generated by comparison between carrier signal and reference signals, and each pulse width is varied corresponding to the amplitude of a sine wave assessed from the middle of the same pulse. MSPWM being considered, the width of the pulse does not change consequently with modulation index variation. first and last 60º intervals per half cycle as shown in Figure 4.

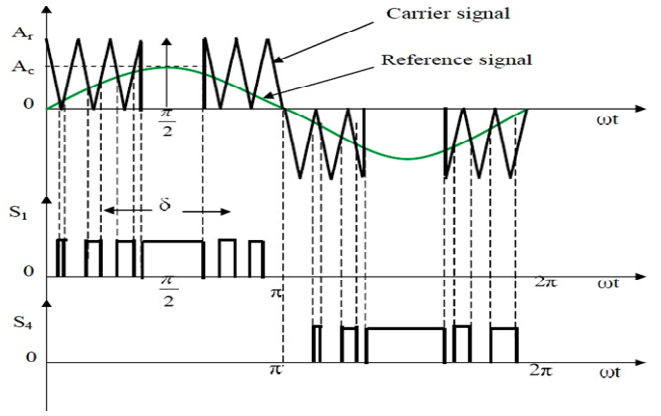


Figure 4: Generation of a Modified Sinusoidal Pulse Width [28, 31 &30].

The improvement of harmonic characteristics results from an increase in fundamental component. This modulation technique has several advantages such as increased fundamental component, improved harmonic characteristics, reduced number of switching power devices and minimum switching losses. Figure 5 show a proposed method by Josias. W Makhubele, *et al* (2004) for controlling output boost of D.C-A.C. inverter with controller modifications.

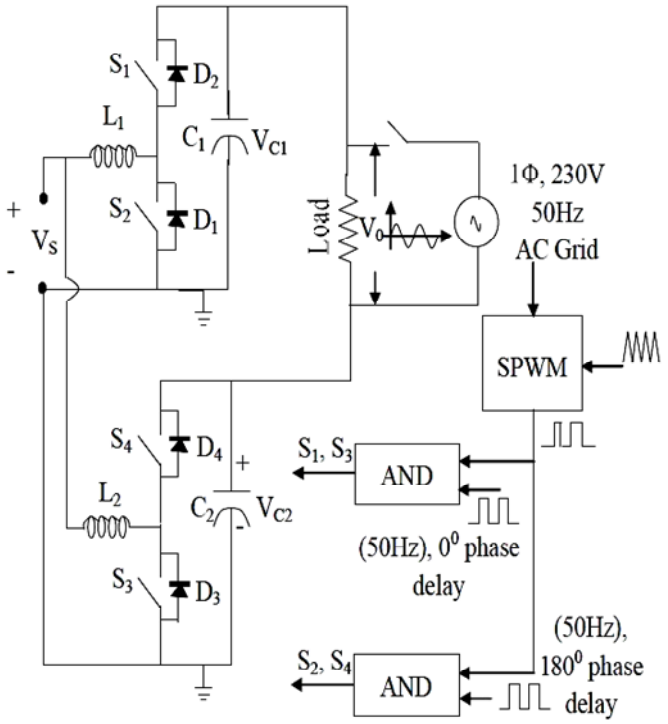


Figure 5: Sinusoidal Pulse Width Modulation Technique-based modified controller of boost inverter [31].

III. Simulation Results

A. Sinusoidal PWM technique

Most inverters which have stable D.C. input voltage usually their magnitude is constant. Triangular or Carrier signal is used as a switching frequency in the inverter. The switching frequency is basically remaining constant. The speed of the inverter switches (turn on/off) times is controlled by this frequency. The switch duty ratio gets modulated by V_{ref} as its primary function. The carrier waveform V_c is at switching frequency F_c. The reference signal, used to modulate the switch duty ratio, and has frequency F_r. The presence of harmonics in the switching frequency affects the inverter output, causing it to have harmonic contents. Inverter switches have a duty cycle that is called Modulation ratio M_a.

In Bipolar SPWM technique, output pulses are generated by a comparison between sinusoidal signal and triangular signal. Output pulses generated in bipolar switching have been shown in Figure 6. $+V_{dc}$ output is generated when sine reference signal amplitude is higher than the triangula signal, and the output is at $-V_{dc}$ when sine reference signal is below or lower than the triangular signal as shown in Figure 6.1 and figure 6.2 respectively [28].

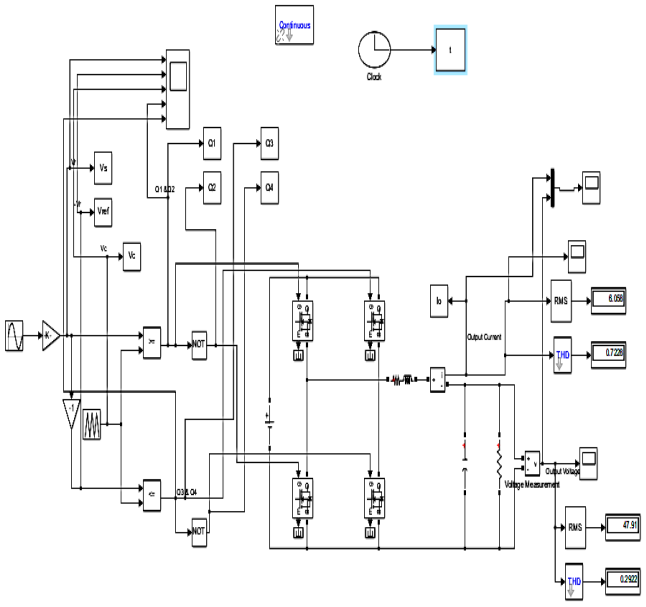


Figure 6: SPWM controller MATLAB Model.

As shown in Figure 6, the following are the parameters used in this model:

The line inductance and resistance are 1mH & 1mΩ, load resistance of 100Ω and capacitance of 1000µF, DC voltage of 100V at a switching frequency of 2kHz and fundamental frequency of 50Hz.

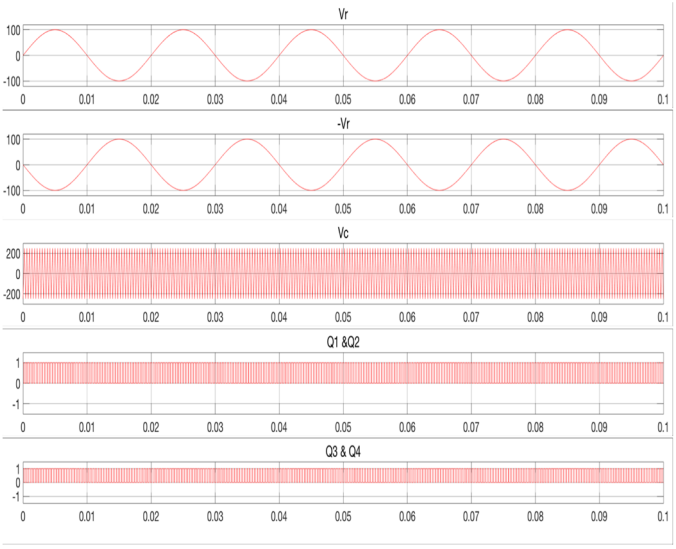


Figure 6.1: Simulated performance of Unipolar PWM inverter (fs=2kHz)

Unipolar SPWM switching scheme that is shown in Figure 6.1 was achieved by comparing two sinusoidal reference waveforms (Vr and -Vr) that have the equal frequency and size but are 180°phase shifted compared to carrier signal (Vc). The output generated by switches Q1 & Q2 and Q2 & Q3 is shown in Figure 6.1. The signal transforms between 0 and 1V of the fundamental frequency.

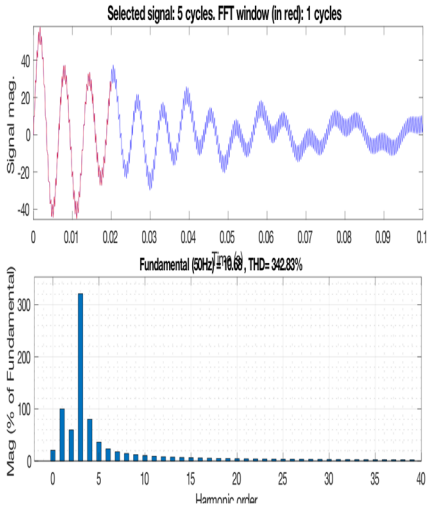


Figure 6.2: FFT results of Unipolar SPWM inverter output current (Io)

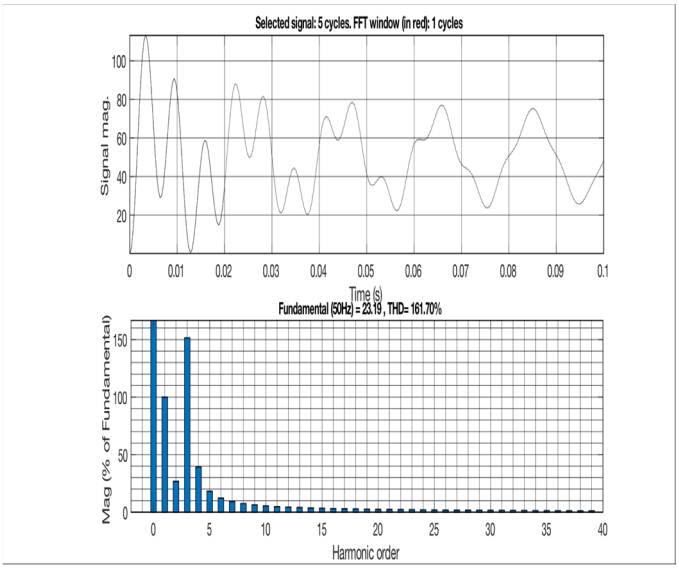


Figure 6.3: FFT results of Unipolar SPWM inverter output voltage (Vo)

All waveform signals shown in Figure 6.2 &6.3 are FFT analysis of both output voltage and current at a switching frequency of 2 kHz. Modulation amplitude is 0.1 at 100V supply. From Figure 6.2, output current signal has high THD compared to output voltage as shown in the figures 6.3. When the gain is decreased the output voltage and current waveforms have less THD and the output waveforms are more sinusoidal.

B. Hysteresis –Band Current Control Pulse Width Modulation (HCCPWM)

Hysteresis-Band Current Control PWM shown in Figure 7, is simply a spontaneous current control feedback method of Pulse Width Modulation, whereby the definite current

continuously follows the command current within a hysteresis band [29]. As shown in Figure 7, the switching transistors Q1&Q4, and Q2&Q3, the input current *Is* is forced to follow a reference shape. The HCC causes the A.C. current vector to trace the fundamental voltage with relatively negligible time and carelessness to line voltage and load criterion alterations. The HCC has a bandwidth with a switching frequency and peak ripple current related. An optimum band serves a purpose of maintaining a balance between a desirable switching loss of an inverter and the harmonic ripple. HCCPWM region can be easily modulated to square wave voltage state through the quasi PWM region. Deserted neutral load creates several distortions in the current waveform and other problems are caused in high performance machine control by phase deviations [3].

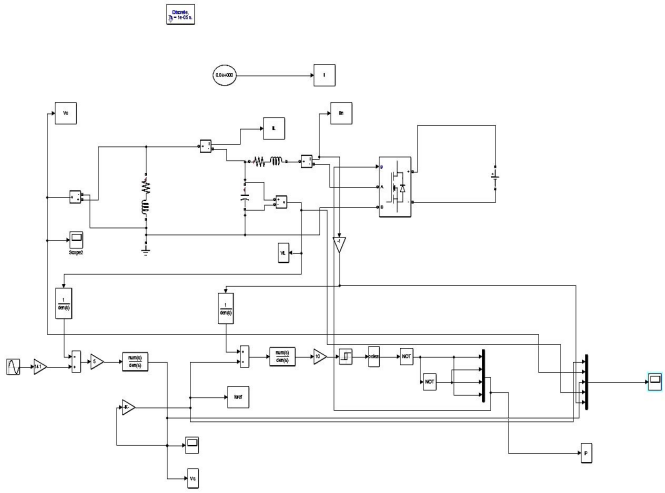


Figure 7: Mat lab HCC Controller

As shown in Figure 7, the following are the parameters used:

The line inductance and resistance are 1mH & 1mΩ, load resistance of 100Ω and capacitance of 1000µF, DC voltage of 100V at a switching frequency of 2kHz and fundamental frequency of 50Hz.

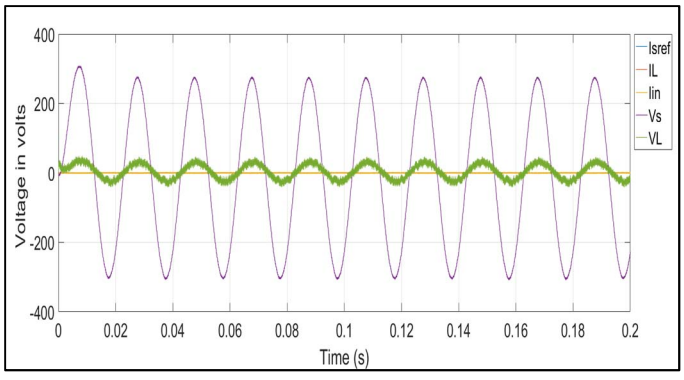


Figure 7.1: HCC waveform signals

Figure 7.1 show voltage and current waveform signals of reference current (Isref), load current (IL), input current (Iin), supply voltage (Vs) and load voltage (VL). Vs switches between zero and positive or zero to negative voltage. Vs is 250v and is sinusoidal in shape. Isref trace the fundamental current with the hysteresis band. VL is the load voltage or output voltage resulting from the load.

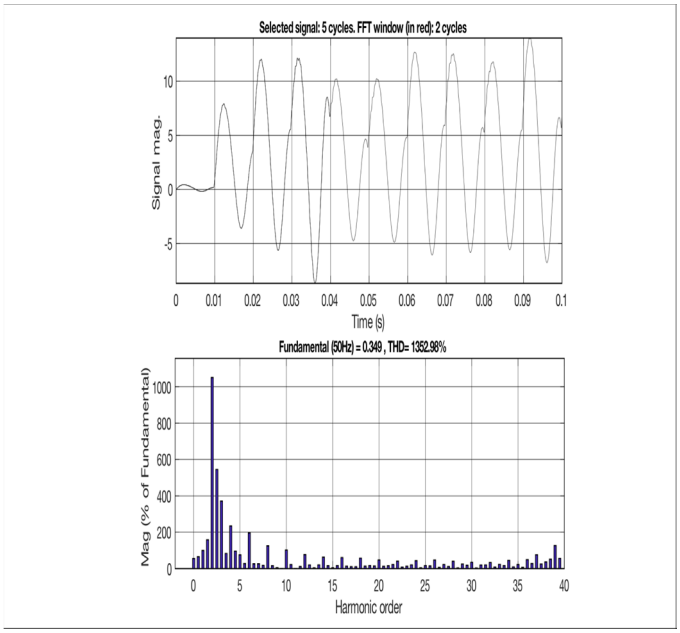


Figure 7.2: Simulated performance of HCC PWM inverter output voltage (Vo) (fs=2kz).

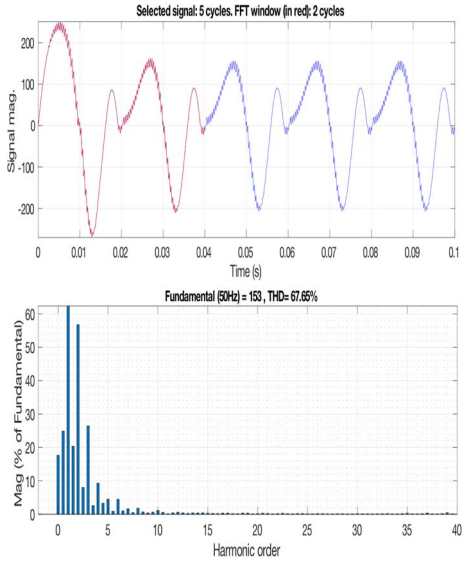


Figure 7.3: Simulated performance of HCC PWM inverter output current (Io) (fs=2kz).

As shown in Figure 7.2 &7.3, FFT analysis of both output voltage and current at a switching frequency of 2 kHz. Modulation amplitude is 0.1 at 100V supply. Voltage and current output have & 1352.90% & 67.65% THD. The higher the gain results, the higher THD value and vice versa.

C. Modified Sinusoidal Pulse Width Modulation(MSPWM)

The Sinusoidal Pulse Width Modulation technique with a modified controller is shown in Figure 8, less research has been done as compared to the standard SPWM. Therefore, the Modified SPWM technique have more prevalence as compared to the SPWM. When MSPWM is used pulse, width remains almost constant with the variation of modulation index due to sine wave attributes. Hence SPWM technique is altered so that during the first and last 60 degrees interval per half cycle, carrier signal can be generated. This technique has several advantages mention in the literature as compared to the traditional SPWM technique

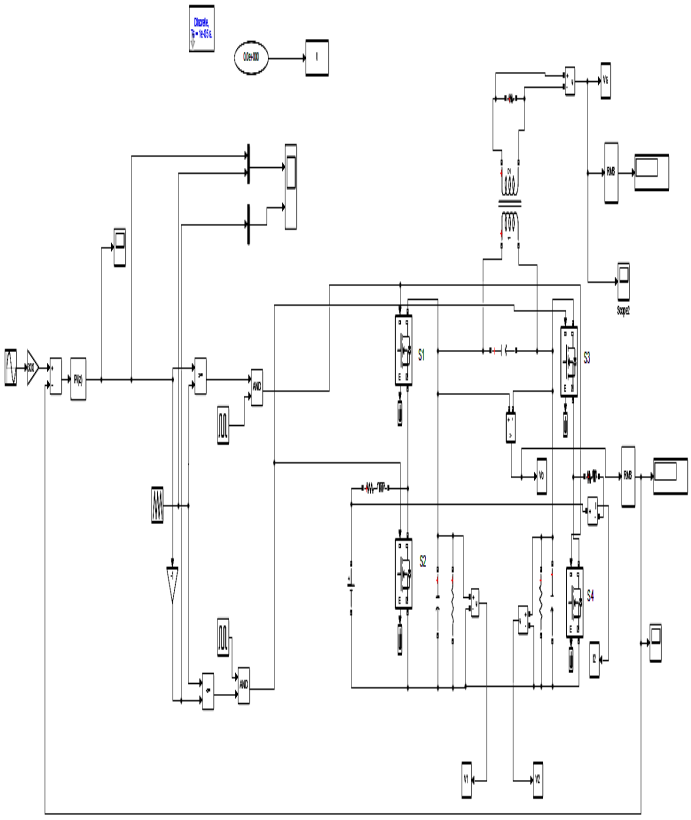


Figure 8: Modified SPWM MATLAB/ Simulink model.

As shown in Figure 8, the following are the parameters used:

The line inductance and resistance are 1mH & 1mΩ, load resistance of 100Ω and capacitance of 1000µF, DC voltage of 100V at a switching frequency of 2kHz and fundamental frequency of 50Hz.

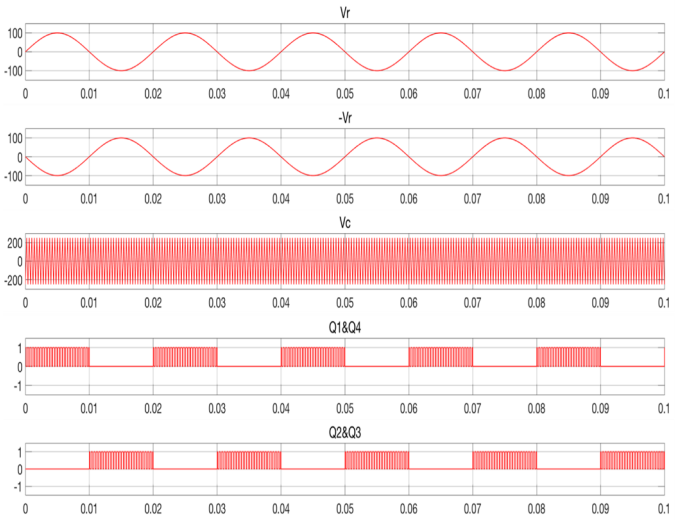


Figure 8.1: Simulated performance of Unipolar PWM inverter (fs=2kHz)

Unipolar SPWM switching scheme that is shown in Figure 8.1 was achieved by comparing two Sinusoidal reference waveforms (Vr and -Vr) that have the same frequency and magnitude but with 180°phase shifted compared with triangular signal (Vc). The output generated by switches Q1 & Q2 and Q2 & Q3 is shown in Figure 8.1. The signal transforms between 0 and 1 V of the fundamental frequency.

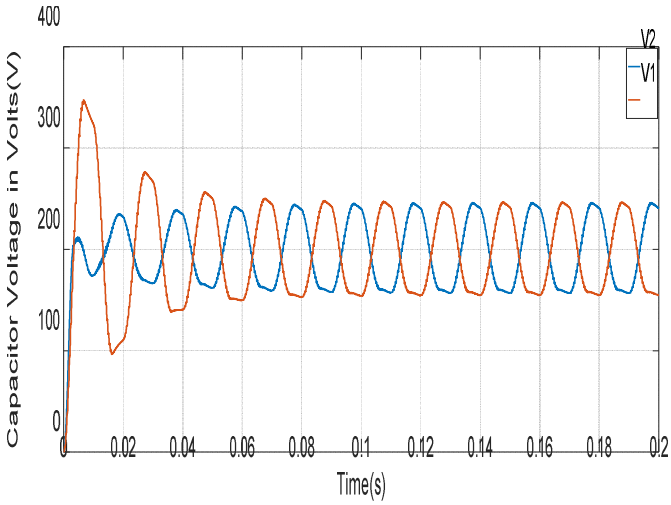


Figure 8.2: Capacitor maximum voltage (V_1 & V_2).

The capacitor voltages (V_1 & V_2) of the boost inverter are kept at constant magnitude for both load and ripple voltage. The phase shift is maintained at 180° between the two capacitor voltages (V_1 & V_2) of the bi-directional boost DC-DC converters. Therefore, the proposed controller controls the capacitor voltages of boost inverter for any alterations in the load of up to a set maximum value.

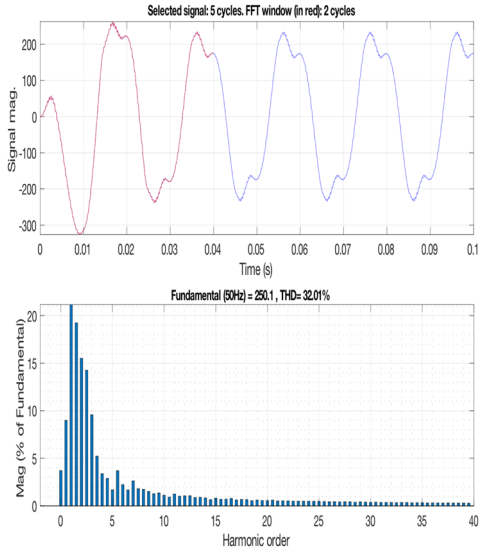


Figure 8.3: FFT results of Unipolar MSPWM inverter output voltage (Vo)

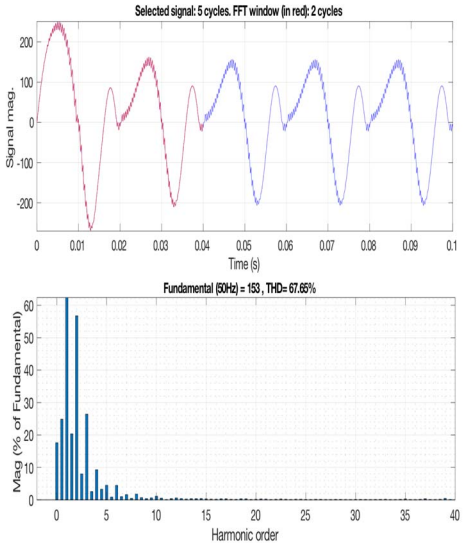


Figure 8.4: FFT results of Unipolar MSPWM inverter output current (Io)

All waveform signals shown in Figure 8.4 &8.5 are FFT analysis of both output voltage and current at a switching frequency of 2 kHz. Modulation amplitude is 0.1 at 100V supply. Current output signal has high THD compared to output voltage as shown in the figures Figure 8.4 &8.5. When the gain is at minimum value the resultant output (Vo &Io), have less THD and more sinusoidal waveform is obtained.

IV. CONCLUSION

MATLAB/Simulink software was used to designed, evaluate and analyze the output voltage and current results between the three most-used modulation techniques. The modulation techniques which have been in use are SPWM, HCCPWM and MSPWM. These PWM techniques are used in multilevel inverters. The main aim of this study is to obtain output voltage and current waveform having a minimum Total Harmonic Distortion (THD). Unipolar switching scheme was used in both SPWM and MSPWM techniques. Results obtained from MATLAB /Simulink were analyzed. Sinusoidal Pulse Width Modulation technique was identified to be the simpler method to implement in a cost-effective manner. SPWM technique have high THD. Output current and voltage of SPWM are 342.83 % & 161.70 % THD. This indicate that the amount of harmonic content is high at both output current and voltage waveform. SPWM scheme has high losses and poor linearity in voltage and current control. HCCPWM produced output current and voltage of 67.65 % & 1352 % THD. In this scheme the switching losses are reduced and the linearity of voltage and current control are better compared to SPWM. HCCPWM has Unstable PWM frequency as its main disadvantage. Modified Sinusoidal PWM technique produces output current and voltage of 32.01% & 67.65 % of THD. This scheme has low losses and voltage and current control is more linear compared to SPWM and HCCPWM. Same parameters were used for all three modulation techniques that were simulated. By comparing the results between SPWM, HCCPWM & MSPWM techniques in a single-phase bridge inverter, the best sinusoidal output signal and low THD is produced by MSPWM technique.

V. REFERENCES

[1] Sandeep Phonate, "Analysis of Single Phase SPWM Inverter", Mewar University, Department of Electrical and Electronics Engineering, Chittorgarh, Rajasthan, India, International Journal of Science and Research (IJSR), ISSN (Online):2319-7064, Impact Factor (2012):3.358

[2] Zina Boussada, Omessaad Elbeji and Mouna Benhamed, "Modelling of Diode Clamped Inverter using SPWM Technique", Photovoltaic Wind and Geothermal Systems Research Unit Electrical department, National Engineering school of Gabes, Zrig 6029 Gabes, Tunisia

[3] Anas Mansoor and Pratibha Tiwari, "Performance Analysis of Single Phase Inverter", (Department of Electrical & Electronics Engineering, Shepherd School of Engineering & Technology, Sam Higginbottom Institute of Agriculture Technology & Sciences Allahabad, India), International Journal of Morden Engineering Research (IJMER),|IJMER|ISSN:2249-6645| Vol. 4| Iss.4| April.2014|

[4] Anuja Namaboodiri and Harshal S. Wani, "Unipolar and Bipolar PWM Inverter", Power Electronics Engineering, Vishwakarma Govt. Engineering College, Chandkheda, International Journal for Innovative Research in Science & Technology| Volume 1|Issue 7| December 2014, ISSN (Online): 2349-6010

[5] Bhavika Shah, S.C. Ramangari and M.M. Renge, "Analytical and Comparative Study of FHI-SPWM and SPWM Control Technique of Five- Phase VSI ", Dept. of Electrical Engineering, SRCOEM, Nagpur, India, IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES-2016)

[6] G. Schettino, V. Castiglia, F. Genduso, P. Livreri, R. Miceli, P. Romano and F. Viola, "Simulation of a Single-Phase Five-Level Cascaded H-Bridge Inverter with Multicarrier SPWM B-Spline Based Modulation Techniques", Department of Energy, Information Engineering and Mathematics Model University of Palermo, Italy, 2017 Twelfth International Conference on Ecological Vehicles and Renewable Energies (EVER).

[7] Bhavika Shah, S.C. Rangari, and M.M. Renge, "Analytical and Comparative Study of FHI-SPWM and SPWM Control Technique of Five -Phase VSI", Department of Electrical Engineering, SRCOEM, Nagpur, India, 1st IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES-2016).

[8] B. Ismail, S. Taib MIEEE, A.R. Mohd Saad, M. Isa, C.M. Hadzer, "Development of a Single Phase SPWM Microcontroller-Based Inverter",

First International Power and Energy Conference PECon 2006 November 28-29, 2006, Putrajaya, Malaysia.
[9] Idkhajine L., Monmasson E., and Maalouf A., Extended Kalman, "Filter for AC drive Sensor less Speed Controller-FPGA-Based solution or DSP-Based solution", Proceeding of the 2010 international Symposium on Industrial electronics, ISIE 2010, July 2010.
[10] Pratik Udakhe, Dipesh Atkar, Sateesh Chiriki, and V.B. Borghate, "Comparison of Different Types of SPWM Techniques for Three Phase Seven Level Cascaded H-Bridge Inverter", EEE Department VNIT, Nagpur 440010, 1st IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES-2016).
[11] Javier Munoz, Jaime Rohten, Jose Espinoza, Pedro Melin, Carlos Baier, and Marco Rivera, "Review of Current Control Techniques for a Cascaded H-Bridge STATCOM", Department of Industrial Technologies, 2015 IEEE.
[12] Ezzidin Hassan Elmabrouk Aboadla, Majdee T., Khairil A. Bin Aznan, et al. "A Comparative Study Between SPWM and SHE-PWM Modulation Techniques for DC-AC Inverters", Department of Electrical and Computer Engineering, KoE, International Islamic University Malaysia.
[13] Hong Li, Hitching Yang, Boyo Wang, Vassillios G. Agelidis and Bo Zhang, "On Thermal Impact of Chaotic Frequency Modulation SPWM Techniques", IEEE transactions on Industrial Electronics, Vol.64, No.3, March 2017.
[14] Meghana N Gujjar and Pradeep Kumar, "Comparative analysis of Field Oriented control of BLDC motor using SPWM and SVPWM techniques", Depart. Of Electrical and Electronics, NMAMIT, Nite, India, 2017 2nd international Conference on recent Trends in Electronics Information & Communication Technology (RTEICT), May 19-20, 2017, India.
[15] R.M. Pindoriya, A.K. Mishra, B.S. Rajpurohit, and R. Kumar, "Performance Analysis of Control Strategies of Permanent Magnet Synchronous Motor", India Institute of Technology Mandi, Himachal Pradesh, India, 2016 IEEE.
[16] K. Vijiyakumar, R. Sundar Raj and S. Kunnam, "Application of Sinusoidal Pulse width Modulation Based Matrix Converter as Revolutionized Power Electronic Converter", Department of EEE, Kalasalinggam University, Krishnankoli 626126, India.
[17] Naim Suleyman Ting, Yusuf Yasa, Ismail Aksoy, and Yakup Sahin, "Comparison of SVPWM, SPWM, and HCC Control Techniques in power Control of PMSG used in Wind Turbine Systems", Department Yildiz Technical University Istanbul, Turkey
[18] Suman Chandra Ghosh and Srinivas Bhaskar Karanki, "PV Supported Unified Power Quality Conditioner Using Space Vector Pulse Width Modulation", India Institute of Technology Bhubaneswar, Arugul, India, 2017 National Power electronics Conference (NPEC) College of Engineering Pune, India, Dec 18-20, 2017.
[19] A.N. Arvindan, "Single -Sensor HCC PWM Control of 3-Phase Boost type AC-DC Converter giving Unity Power factor", Department of Electrical and Electronic Engineering, SSN College of Engineering, Anna University, Chennai, India, 2012 IEEE
[20] Alireza Fereidouni, and Mohammed A.S. Masoum, "A new Space Vector Hysteresis Current Control Modulation for Active Power Filter Applications", Department of Electrical and Computer, Curtin University, Perth, WA, Australia
[21] Arindam Mandal, V. Inder Kumar, and Santanu Kapat. "Multi-Band Mixed-signal Hysteresis Current Control for EMI Reduction in Switch-Mode Power Supplies", Department of Electrical Engineering India Institute of technology Kharagpur, West Bengal, India,2016 IEEE
[22] Deepak Ronanki, Najath A. Azeez, Lalit Patnaik, and Sheldon S. Williamson, "Hybrid Multi-Carrier PWM technique with Computationally Efficient Voltage Balancing Algorithm for Modular Multilevel Converter", Smart Transportation Electrification and Energy Research (STEER) Group, Department of Electrical, Computer, and Software Engineering University of Ontario Institute of Technology (UOIT), Oshawa Canada.
[23] Ezzidin Hassan Elmabrouk Aboadla, Majdee T., et al, "A Comparative Study Between SPWM and SHE-PWM modulation Techniques for DC-AC Inverters", Department of Electrical and Computer Engineering, K0E, International Islamic University Malaysia (IIUM)
[24] Ezzidin Hassan Elmabrouk Aboadla, Sheroz Khan et al, "Effect of Modulation Index of Pulse Width Modulation Index of Pulse Width Modulation Inverter on Total Harmonic Distortion for Sinusoidal", Department of Electrical and Computer Engineering, Kulliyyah of Engineering, International Islamic University of Malaysia (IIUM)
[25] Sachin Maheshri and Prabodh Khampariya, "Simulation of Single phase SPWM (Unipolar) inverter", S.S.S.I.T., Sehore MP, international journal of innovative research in advanced Engineering (IJIRAE), Volume 1, Issue 9 (October 2014)
[26] B. Venkata Ranganadh, A. Mallikarjuna Prasad, and Madichetty Sreedhar, "Modelling and Simulation of a Hysteresis Band Pulse Width Modulated Current Controller Applied To a Three Phase Voltage Source Inverter by using Matlab", PG Scholar, Department of Electronics & communication Engineering, JNTUK, Kakinada, India, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol.2, Issue 9, September 2013
[27] Narri Yadaiah, N. Ravi Kumar, and G. Tulasi Ram Das, "Comparison of Conventional, Fuzzy Logic and Neutral Based HBPWM Current Controllers", 2006 International Joint Conference on Neural Networks Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada, July 16-21, 2006
[28] Kalpesh R. Chaudhary, MJ. Mod2, "Switching Control of Inverter Using SPWM Technique", Instrument and Control Management LDCE, Ahmedabad, 2015 IJRT| Volume 1 Issue 12| ISSN:2349-6002
[29] D.V. Nicolae: "Consideration of the most appropriate Techniques for the control of the multi-purpose (reversible) switching-mode converters", MSc (Eng.) Simulation report, Technikon Northern Gauteng.
[30] K. Georgakas and A. Safacas, "Modified Sinusoidal Pulse Width Modulation Operation Technique of an AC-AC Single Phase Converter to Optimise the Power Factor". June 2010.IET Power Electronics 3 (3): 454-464: available on https://www.researchgate.net/publication/224134515
[31] Shodhganga.inflibnet.ac.in/bitstream/1060/33606/10/10_chapter 5.pdf

Call Data Record Based Recommender Systems for Mobile Subscribers

Elias Mbongeni Sibanda and Tranos Zuva
Department of information and Communication Technology
Vaal University of Technology
Vanderbijlpark, South Africa
mesibanda@gmail.com, zuvat@vut.ac.za

***Abstract*—Recommendation systems for mobile phones are of great importance for mobile operators to achieve their desired profit targets. In a client inferred market, the number of contract users and contract phones is especially significant for mobile service operators. The tremendous growth in the number of available mobile cellular telephone contracts necessitates the need for a recommender system to assist users discover suitable contracts based on their usage patterns. This study used a hybrid of both collaborative and content-based filtering. A prototype of a mobile recommender system was developed and evaluated using precision and recall. The developed recommender system was able to successfully recommend packages to subscribers. A precision-recall curve was produced, and it showed good performance of the system. This study successfully showed that a hybrid system was able to recommend products to the mobile subscribers.**

Keywords—mobile recommender systems, mobile subscriber, collaborative filtering

I. INTRODUCTION

Recommender systems automate the process of recommending products, services or information items (hereinafter referred to as items) to consumers (referred to as subscribers) based on several types of data concerning users, items and previous interactions between users and items [1]. There seems to be no single algorithm that is best for this purpose and relative performances of different algorithms are largely domain and data-dependent [2].

This study reports on the development of a mobile recommendation system to assist TELecommunications COmpanies (TELCOs) in making product recommendations. Recommender systems are a subclass of what rating a user may give to a certain item. They are widely used in the movie, shopping and other domains. The primary purpose of Recommender systems' is to pass a recommendation of items to a user, the items have the likelihood that the user may not be interested in them.

In a web-based environment, recommender systems recommend items to a user based on item ratings and similarity. Recent studies have shown that using a Recommender System (RS) can prime to amplified sales dimensions in the short-term and in the long-term or aid to growth in sales by pointing users to further parts of the product catalog [3]. This technology has been successful for web users in providing targeted item recommendations but only a few have been designed for mobile users [4].

Recommendation systems use several different technologies. These items can be classified to two wide-ranging groups. Content-based (CB) systems which scrutinize details of the recommended items and Collaborative filtering (CF) systems which recommend items on the basis of similarity amongst items and or users. In a recommender system program, the classes of entities are referred to as items and users. Users have inclinations for specific items, and these inclinations must be figured out of the data [5].

For a recommendation to be possible, the researcher proposes that a user inputs his/her mobile number and the recommender system then recommends all suitable items based on information retrieved from the database. It is important that when recommendations are made from a mobile subscriber recommender system, the recommendation is based on Call Data Records (CDR) for that subscriber. The use of CDR will enable the recommender system to produce valuable results because the information retrieved is factual and the user is most likely to react to the recommendation. For the purpose that the recommendation will be personal (based on user data) and not general (based on a group information, this approach will eliminate erroneous recommendations of products that the user cannot afford.

Dealing with the cold-start problem for new items has been one of the longest-standing challenges in recommender systems [6]. Instead of submissively consuming recommended items, users regularly need to govern their interactions with the items that are in the system. The recommendations are frequently best utilized as aides through a mind boggling item space, giving the user a chance to pick in what direction to move at each progression [7].

The primary usage of the first mobile devices was voice and for decades TELCOs generated their revenue from voice. However, this has changed with the availability of the internet and the introduction of instant messaging platforms like Facebook, WhatsApp, Imo etc. In the early 1980's there was very little use for mobile recommender systems due to the limited services and capability of mobile devices.

A. Statement of the Problem

Current mobile telecom networks store large amounts of data including the duration of the call, the time the call was executed, the subscriber payment plan (Pre-paid or Post-paid) etc. in their databases and this information is not used for recommendations. There is an ever-increasing complexity of understanding behavioral patterns of subscribers' preferences which render the traditional recommendation approaches not competent in satisfying customer needs. A new recommendation paradigm is required that uses techniques that exploit user-specific information based on time, usage, interests, etc., instead of common methods to generate recommendations.

One of the greatest challenges faced by recommender systems is the lack of recommendation accuracy. It is evident from [8] that this is caused by the ambiguity in querying of information and or the structure of the database. Adding to the challenge is the gaining popularity of different kinds of devices that use different communication standards. This has made it difficult for mobile subscriber recommender systems to be accurate in recommending items to users, leading to recommendation errors. A user might not have a device that is internet enabled and that user then receives a recommendation that has data bundles as part of the package. That kind of recommendation does not address the user's needs and the user is likely not to react to the recommendation.

B. Research Question

How can a mobile subscriber recommender system be developed in a way that it takes into consideration usage patterns of the mobile subscriber to make a recommendation?

This study intends to answer the below sub-questions:

- What recommender systems methods and techniques are currently available?
- How can a mobile subscriber recommender system be developed in a way that recommends usable products to a mobile subscriber (user)?
- How to measure the efficiency of the developed recommender system?

C. Objectives of the study

- To compare different recommender systems algorithms from the literature.
- To highlight the challenges in recommender systems.
- To measure the effectiveness of the proposed recommender system.

D. significance of the study

The use of CDR based recommendation is novel and little research has been done in this area, considering that most recommender systems are based on user ratings and user history. However, this study is expected to make a major contribution towards how TELCOs recommend products to users. This study is expected to also contribute in highlighting the problems, challenges and other issues that are faced by current recommender systems for mobile subscribers. The perception of how mobile subscriber recommender systems are developed in the future is expected to be changed.

This study is arranged as follows section II related work, section III methodology, section IV Evaluation metrics, section V Evaluation and results, Section VI conclusion and section VII Future work.

II. RELATED WORK

Mobile devices are gradually increasing in popularity since the prices for these devices are gradually decreasing and people can afford to own them [9]. It is clear from [10] that the use of cellular phones as well as the hasty progress of the internet has caused an information overload issue. However, in a mobile recommender system space the usage of mobile devices brings about a new phenomenon and changes the focus of recommendations to individuals, thus making personalized recommendations an important factor. It is evident from [11] that in a mobile situation, information personalization is more difficult for the reason that there are constrains in mobile devices regarding displays and bandwidth etc. However, it can be said that personalization is not a new research topic within the recommender systems theory because Amazon.com, has been doing it successfully [10].

The groundbreaking technology, mobile computing allows access to information instantly, and it is not location dependent. Amid others, there has been studies in the tourism field aimed at the use of mobile technology. Along this line, mobile RSs represent a relatively recent thread of research with numerous potential application fields (e.g., mobile shopping, advertising and content provisioning) [7]. For example, [12] proposed an area mindful recommender framework that suits clients' shopping needs with area subordinate seller promotions, it likewise empowers the production of customized campaigns focusing on users as indicated by their area, needs and gadgets profiles.

A. Recommender system algorithms

Numerous studies have been directed on the arrangement of context-aware information services like customized news conveyance [13, 14].Mobile information recommendation is being very prevalent due to the increasing range, convenience and use of mobile information services [13]. Collaborative filtering methods encompasses probabilistic and non-probabilistic methods. It can be said that non-probabilistic replicas are expansively used by experts and Probabilistic models have been widely used in the machine learning space [15].

Non-probabilistic algorithms: it is evident from [15] that the most well-known CF algorithms are nearest neighbor algorithms. There are also graph-based algorithms [16], neural networks [17], and rule-mining algorithms [18].

Neighborhood algorithms are computationally expensive. For example, to accurately get the distance between two neighbors the whole table must be scanned, and the distance must be calculated for different items.

According to [19] Collaborative filtering detects users (neighbors) with similar interests to those of a specific customer and recommends products neighbors of a given customer. However, regardless of their success their extensive use has exposed two major limitations [20]. The first is related to sparsity. The number of ratings already obtained is very small compared to the number of ratings that need to be predicted for the reason that typical collaborative filtering requires clear non-binary user ratings for comparable products. As a result, collaborative-filtering based recommendations cannot accurately compute the neighborhood and identify the products to recommend.

1.1.1 Similarity in recommender systems

Below is a representation of three (3) types of similarities that are applicable in recommender systems depending on the specific objective. These three are, namely, cosine-based similarity, adjusted cosine-based similarity and correlation-based similarity.

a) Cosine-based similarity

In cosine-based similarity two items are assumed as two vectors in the m dimensional user-space. The similarity between them is calculated by computing the cosine of the angle between the two vectors. Formally, in the $m*n$ ratings matrix below the similarity between items i and j, denoted by $sim(i,j)$ is given by:

$$sim(i,j) = \cos(\bar{i},\bar{j}) = \frac{\bar{i}.\bar{j}}{\|\bar{i}\|_2 * \|\bar{j}\|_2}$$

where "." Denotes the dot-product of the two vectors and *sim (i, j)* is the similarity between item *i* and item *j*.

b) Correlation-based similarity

The similarity between two items i and j is measured by computing the $Pearson-r$ correlation $corr_{i,j}$. To make the correlation computation accurate, co-rated cases (i.e., cases where the users rated both i and j) must be. The set of users who both rated i and j are denoted by U then the correlation similarity is given by:

$$sim(i,j) = \frac{\sum_{u\in U}(R_{u,i}-\overline{R_i})(R_{u,j}-\overline{R_j})}{\sqrt{\sum_{u\in U}(R_{u,i}-\overline{R_i})^2}\sqrt{(R_{u,j}-\overline{R_j})^2}}$$

where $R_{u,i}$ denotes the rating of user u on item *i*, R is the average rating of the i-th item and $R_{u,j}$ denotes the rating of user u on item j, R_j is the average rating of the j-th item

c) Adjusted cosine similarity

One vital difference between the similarity computation in user-based CF and item-based CF is that in case of user-based CF the similarity is computed along the rows of the matrix but in case of the item-based CF, the similarity is computed along the columns, i.e. each pair in the co-rated set corresponds to a different user. Computing similarity using basic cosine measure in item-based scenarios has one vital downside the differences in rating scale between different users are not considered. The adjusted cosine similarity offsets this drawback by subtracting the corresponding user average from each co-rated pair. Formally, the similarity between items i and j using this scheme is given by:

$$sim(i,j) = \frac{\sum_{u\in U}(R_{u,i}-\overline{R_i})(R_{u,j}-\overline{R_j})}{\sqrt{\sum_{u\in U}(R_{u,i}-\overline{R_j})^2}\sqrt{(R_{u,j}-\overline{R_j})^2}}$$

where $\overline{R}u$ is the average of the $u-th$ user's ratings and $\overline{R_j}$ is the mean average rating given by user j.

B. Hybridazation

Hybridization is the amalgamation, otherwise the use of all or some of the features from diverse algorithms to generate a recommendation. While each of the algorithms described above is dominant in their respective fields, hybridization is in need and it is gaining attention [21, 22]. Moreover hybrid algorithms have been categorized according to their design [23].

Hybridization is not a novel concept in recommender systems, but it has been applied differently in theory. According to [11] a hybrid recommender system is a combination of different recommender systems to improve the retrieval of information. Different combinations can be fused together to achieve a hybrid for instance: weighted, switching, mixed, feature combination or augmentation or cascading [24]. More information on hybrid algorithms can be found in [23]

C. Issues affecting mobile reccomender systems

Most of the issues that are highlighted when designing recommender systems for mobile users refer to mobile phones and sim-card enabled tablets. It is imperative to note that recommendations on tiny screen devices can be problematic and annoying for end-users. It is evident that users are able to read and understand information on small interfaces; however the screen size has a performance impact [25]. In addition, a small screen user is less effective in completing assigned tasks as compared to large screens users.

Another issue associated with mobile recommender systems is data costs. Most mobile recommender systems are run online, and a user must have data to interact. The cost of data often impacts how much time a user will spend on a site. Due to the limitations stated above, information search browsing, and in particular item recommendations, is a problem on mobile devices. In addition, a user on a small

screen is less effective in completing an assigned task when compared to users with large screens.

III. METHODOLOGY

This study literature exploration and experiments were conducted. The data was collected from the OPCO. The data was then loaded into a Java data manipulation tool, pre-Processed, analyzed, and interpreted by the researcher. The pre-processed data was thus used in experiments to produce recommendations to the subscriber. This study follows the steps as illustrated in Figure 1. The data was collected, then converted into the correct format, after which it was pre-processed. After completion of the above steps the data was then classified, and prediction took place. The data was then loaded into the recommender system and the recommendations computed. After recommending to the subscriber the last steps involve evaluating the system and analyzing the results.

Fig. 1. Illustration of how the methodology flows

In this study, two separate sets of recommendations were generated and merged them together to develop a final set of recommendations.

A. COLLABORATIVE FILTERING IN SUBSCRIBER RECOMMENDER SYSTEMS

Mobile subscriber data was collected from the OPCO and used in experiments to predict the future usage patterns and recommendations. The system performance results were evaluated using precision and recall. A comparison was done between the recommendation generated and the usage patterns of the subscribers.

1) Subscriber *based nearest neighbour*

In subscriber based nearest neighbor collaborative filtering algorithm, the prediction of a possible contract is based on the usage patterns of the subscriber in comparison to the contract users. There must be similarity between the two subscribers for a recommendation to be possible. The crucial part of subscriber nearest neighbor will be that the subscriber is to identify the nearest neighbor of the target subscriber. The formula for Pearson correlation coefficient similarity is given in Equation 4.

$$Subscribersim(u_t,u)=\frac{\sum_{i\varepsilon I_{u,ut}}(R_{u,i}-\bar{R}_u)(R_{u_t,i}-\bar{R}_{u_t})}{\sqrt{(R_{u,i}-\bar{R}_u)^2}\sqrt{\sum_{i\varepsilon I_{u,ut}}(R_{u_t,i}-\bar{R}_{u_t})^2}}$$

where $Subscriber\ sim(u_t,u)$ is a representation of the similarity between subscriber u and u_t, $I_{uu_t}=I(u)\cap I(u_t)$ means the product is used by both the subscribers and u , u_t , and $R_{u,i}$ are the usage totals for product i used by subscriber u and u_t respectively, R_u and R_{u_t} represents the average usage amounts over a time of users u and u_t, respectively.

2) Product nearest neighbour

Product based nearest neighbor algorithms works the same as subscriber based nearest neighbor; however, the difference lies significantly on the product. The recommendation is made purely based on the product nearest neighbor. The similarity is calculated based on the product that is common between two different subscribers. The formula for adjusted based cosine which is mostly used is given in equation 29 below:

$$productsim\ (i_t,j)=\frac{\sum_{u\varepsilon U_{t,j}}(R_{u,i}-\bar{R}_u)(R_{u_t,j}-\bar{R}_{u_t})}{\sqrt{\sum_{u\varepsilon U_{t,j}}(R_{u,i}-\bar{R}_u)^2}\sqrt{\sum_{u\varepsilon U_{t,j}}(R_{u,j}-\bar{R}_u)^2}}$$

where R_u and $R_{u,j}$ represents the usage amounts of subscriber u on product j. $\bar{R}_u$ is the mean of the u^{th} subscriber's usage amount over a certain time and $u_{i,j}$ represents all the subscribers that have used product i and j. The prediction for product based nearest neighbor algorithm for subscribers is carried out using equation 30 below.

$$P_{\substack{(u,j)\\ product-based}} = \frac{\sum_{i \varepsilon R_{ut}} prodsim(i,j) * R_{ut,j}}{\sum_{i \varepsilon R_{ut}} prodsim(i,j)}$$

Where $prodsim(i,j)$ is the product similarity between product i and product j and $R_{ut,j}$ is the usage amount of user *ut* on product *j*.

B. comparison of Collaborative Filtering and Content Based Filtering

Both collaborative filtering and content-based techniques have their own strengths and weaknesses. The combination of the two techniques allows the strength of the one technique to overcome the weakness of the other technique; hence they are usually combined, and a hybrid is formed.

TABLE I. *COMPARISON BETWEEN COLLABORATIVE FILTERING AND CONTENT BASED*

Focus area	Collaborative filtering	Content based
No need for domain knowledge	Yes	Yes
Adaptive	Yes	Yes
Quality improves as new data is loaded	Yes	Yes
Implicit feedback adequacy	Yes	Yes
Impacted by cold start problem	Yes	No
Susceptible to new subscriber problem	Yes	Yes
Susceptible to new product problem	No	Yes
Quality of the recommendation depends on the size of the data	Yes	Yes

Both CF and Content Based Filtering (CBF) techniques have no need for domain knowledge. This means that content based and collaborative filtering are of adaptive nature and can build recommendations with implicit feedback and no domain knowledge [26]. However, both techniques improve as there is more data. CF is susceptible to new products which is a problem that CBF addresses. The more data there is on the system the better the quality of the recommendation.

IV. EVALUATION METRICS

There are several evaluation methods that can be applied and used in recommender systems. These methods differ depending on the goal that the researcher wishes to achieve. The evaluation of this study is two-fold: the first part of the evaluation is offline, on the server side and the other part is done online and on the client side. It is imperative to note that the recommender system in this study can be seen as a client server architecture whereby the client side is the subscriber mobile device and the server side is equipped with the database and recommendation system used a graphical evaluation precision and recall in evaluating the system from the client side has been used.

To effectively evaluate the system precision and recall were calculated as measures of effectiveness. In this research, precision is defined as the total number of all relevant products divided by the total number of all retrieved products. Precision and recall are represented as follows.

$$Precision = \frac{A}{A+C}$$

$$Recall = \frac{A}{A+B}$$

where A is the number of relevant items recommended, C is the number of relevant items in the database and B is the number of all irrelevant items recommended.

To the subscriber the scalar value of recall indicates the ability of the system to find relevant items as per query from the collection of different products. Precision demonstrates the capacity to yield pertinent products according to the query. When all is completed, the subscriber is keen on the pertinent prescribed products. Thus, these measures of precision and recall become measures of interest to the subscribers. The lower the value the more undesirable is the performance of the system. The Precision-Recall graph gives a visual performance of the recommender system and the graph is given in the next section. The system was measured on these measures.

V. RESULTS AND ANALYSIS

Figure 2. Shows that these usage streams were further classified into different hybrid systems and different combinations are revealed within the data. Since not all subscribers are using the same products, further classifications and combinations are possible. Therefore, the combinations have been broken down into different hybrids. It was discovered that the VOICE&SMS hybrid has 59% usage, while DATA&VOICE has a twenty-six percent (26%) usage. Only 2% used DATA&SMS hybrid. A graphical representation of Table 1 can be found in Figure 2.

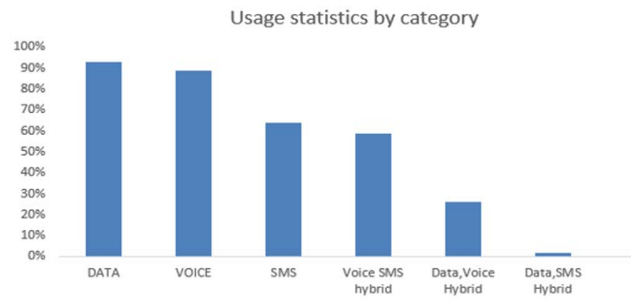


Fig. 2. Results for usage statistics and hybrids

SYSTEM EVALUATION

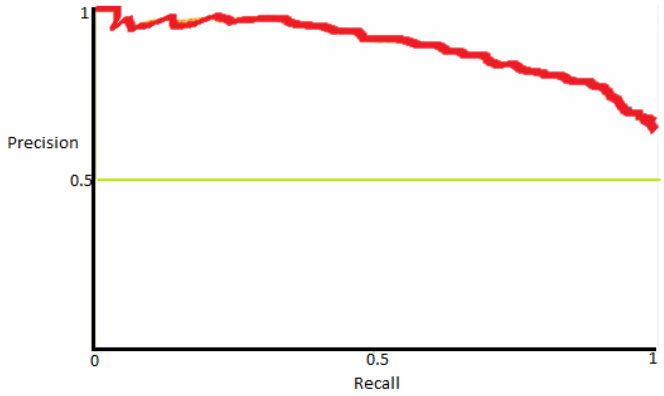


Fig. 3. Precision and recall curve of a binary classifier for correctly predicting the products.

The nature of the precision and recall in this study is for a binary classifier whereby the product recommended can either be correct or not correct. If the recommended product matches the usage, then it is classified as correct and when the product does not match the usage it is classified as incorrect. The precision-recall performance in Figure 3 is exceptional, as the precision increases the recall decreases.

[7] in their study found the precision of their system to be performing around 93%. [27] in their study the system was performing at 75% precision. Therefore, the precision-recall curve represented in figure 3 show good performance of the system with precision of 98%.

VI. CONCLUSION

This study successfully implemented a hybrid recommender system algorithm (content based and collaborative filtering approaches) that was able to recommend correct products to mobile subscribers. A precision-recall curve was produced, and it showed good performance of the system. The study successfully showed that a hybrid system used is capable of recommending products liked by the mobile subscribers.

VII. FUTURE WORK

Future research on this topic involves making the recommender system a location-based by using subscriber location that can be retrieved from the CDR and including more components to the system. These components can be the mobile dealers as well as different handset models nearest the dealer. Further research can also include how the recommender system can be made proactive instead of query based or reactive. In a proactive recommendation system scenario, the recommender will proactively push the recommendations to the subscriber.

REFERENCES

[1] S. Trewin, "Knowledge-based recommender systems," *Encyclopedia of library and information science,* vol. 69, no. Supplement 32, pp. 180, 2000.

[2] A. Deshpande, M. Puerta, M. C. Melguizo, and L. Boves, "A proactive recommendation system for writing: helping without disrupting."

[3] D. Fleder , and K. Hosanagar, "The Impact of Recommender Systems on Sales Diversity," *Blockbuster Culture's Next Rise or Fall,* vol. 55, no. 5, pp. 697-712, 2009.

[4] F. Ricci, and Q. N. Nguyen, "Acquiring and revising preferences in a critique-based mobile recommender system," *IEEE Intelligent Systems* pp. 22–29, 2007.

[5] Y. Koren, R. M. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *IEEE Computer*, no. 8, pp. 30–37, 2009.

[6] K. Riedl, and J. Riedl, "Recommender systems for fast paced users," *A Peaceful Passing*, 2013.

[7] F. Ricci, "Mobile recommender systems," *International Journal of Information Technology and Tourism* vol. 12, no. 3, 2011.

[8] F. Ricci, D. Cavada, N. Mirzadeh, and A. Venturini, "Case-based travel recommendations," no. 12, pp. 50-55, 2006.

[9] J. E. Smith, and A. M. Brown, "Building a culture of learning design: Reconsidering the place of online learning in the tertiary curriculum," 2005.

[10] N. Polatidis, and C. Georgiadis, "Mobile Recommender Systems: An Overview of Technologies and Challenges," *In Informatics and Applications (ICIA)*, 2013.

[11] W. Woerndl, C. Schueller, and R. Wojtech, "A Hybrid Recommender System for Context-aware Recommendations of Mobile Applications," 2007.

[12] W. Yuan, D. Guan, Y.-K. Lee, S. Lee, and S. J. Hur, "Improved trust-aware recommender system using small-worldness of trust networks," *Knowledge-Based Systems,* vol. 23, no. 3, pp. 232-238, 2010.

[13] K. F. Yeung, Y. Yang, and D. Ndzi, "A proactive personalised mobile recommendation system using analytic hierarchy process and Bayesian network," *Journal of Internet Services and Applications,* vol. 3, no. 2, pp. 195-214, September 01, 2012.

[14] C.-Y. Wang, Y.-H. Wu, and S.-C. T. Chou, "Toward a ubiquitous personalized daily-life activity recommendation service with contextual information: a services science perspective," *Information Systems and E-Business Management,* vol. 8, no. 1, pp. 13, 2010.

[15] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen, "Collaborative Filtering Recommender Systems," *The Adaptive Web*, 2007.

[16] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-Based Collaborative Filtering Recommendation Algorithms," *ACM,* vol. 1, 2001.

[17] B. Shams, and S. Haratizadeh, "Graph-based Collaborative Ranking," 2016.

[18] W. LIN, S. ALVAREZ, and C. RUIZ, "Efficient Adaptive{Support Association Rule Mining for Recommender Systems," 2001].

[19] Y. H. Cho, J. K. Kim, and S. H. Kim, "A personalized recommender system based on web usage mining and decision tree induction," pp. 329–342, 2002.

[20] A. L. Godfrey, *A product segmentation approach and its relationship to customer segmentation approaches and*

recommendation system approaches: The University of Texas at Austin, 2007.

[21] M. A. Ghazanfar, and A. Prugel-Bennett, "A scalable, accurate hybrid recommender system." pp. 94-98.

[22] P. Massa, and P. Avesani, "Trust metrics in recommender systems," *Computing with social trust*, pp. 259-285, 2009.

[23] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, *Recommender systems: an introduction*: Cambridge University Press, 2010.

[24] R. Burke, "Hybrid recommender systems: Survey and experiments," *User modeling and user-adapted interaction,* vol. 12, no. 4, pp. 331-370, 2002.

[25] M. Jones, and G. Marsden, *Mobile Interaction Design*: Wiley, 2005.

[26] B. Kumar, and N. Sharma, "Approaches, Issues and Challenges in Recommender Systems: A Systematic Review," *Indian Journal of Science and Technology,* vol. 9, no. 47, 2016.

[27] L. Baltrunas, B. Ludwig, S. Peer, and F. Ricci, "Context relevance assessment and exploitation in mobile recommender systems," *Personal and Ubiquitous Computing,* vol. 16, no. 5, pp. 507-526, 2012.

Integrating a Security Operations Centre with an Organization's Existing Procedures, Policies and Information Technology Systems

Muyowa Mutemwa
Department of Peace, Safety and Security
The Council of Scientic and Industrial Research
Pretoria, South Africa
mmutemwa@csir.co.za

Jabu Mtsweni
Department of Peace, Safety and Security
The Council of Scientic and Industrial Research
Pretoria, South Africa
jmtsweni @csir.co.za

Lukhanyo Zimba
Department of Peace, Safety and Security
The Council of Scientic and Industrial Research
Pretoria, South Africa
lzimba@csir.co.za

***Abstract*—A Cybersecurity Operation Centre (SOC) is a centralized hub for network event monitoring and incident response. SOCs are critical when determining an organization's cybersecurity posture because they can be used to detect, analyze and report on various malicious activities. For most organizations, a SOC is not part of the initial design and implementation of the Information Technology (IT) environment but rather an afterthought. As a result, it is not natively a plug and play component therefore there are integration challenges when a SOC is introduced into an organization. A SOC is an independent hub that needs to be integrated with existing procedures, policies and IT systems of an organization such as the service desk, ticket logging system, reporting, etc. This paper discussed the challenges of integrating a newly developed SOC to an organization's existing IT environment. Firstly, the paper begins by looking at what data sources should be incorporated into the Security Information and Event Management (SIEM) such as which host machines, servers, network end points, software, applications, webservers, etc. for security posture monitoring. That is, which systems need to be monitored first and the order by which the rest of the systems follow. Secondly the paper also describes how to integrate the organization's ticket logging system with the SOC SIEM. That is how the cybersecurity related incidents should be logged by both analysts and non-technical employees of an organization. Also, the priority matrix for incident types and notifications of incidents. Thirdly the paper looks at how to communicate awareness campaigns from the SOC and also how to report on incidents that are found inside the SOC. Lastly the paper looks at how to show value for the large investments that are poured into designing, building and running an SOC.**

Keywords—Cybersecurity Operation Centre, incident response, priority matrix, procedures and policies

I. INTRODUCTION

According to the latest Verizon Data Breach report released during the first quarter of 2018 [1], there were 53 000 incidents and 2 216 confirmed data breaches as cyberattackers hit organisations from small businesses to large enterprises. The loss of customer data could lead to reputation loss or regulatory fines. For an organisation to defend against cybercrimes and cyberattackers, that organisation would need to be able to be aware of potentials threats, detect breaches and events, and respond as quickly as possible to incidents. A Security Operations Centre (SOC) is a centralized monitoring and responding hub consisting of people, process and technologies geared towards creating defences, detecting and responding to incidents. Organization that do not have a SOC and that do not conduct periodical security assessments do not have an understanding of their true cybersecurity posture. That is, they do not know their vulnerabilities neither are they aware when their systems have been breached. In this paper a cybersecurity breach is an unauthorized access to an Information Technology (IT) system.

In this paper a cybersecurity event is when there is a possible hardware or software breach into an IT system. After investigations, an event can either become a false positive, false negative or true negative. False positive is when a security alert was created but there was no malicious activity. False negative is when a security alert was not created but there was malicious activity. True positive is when a security alert was created and there was malicious activity.

In this paper a cybersecurity incident means that there was malicious activity. After investigations, an incident can either be a false negative or true positive. That is there was malicious activity but the alert systems might or might not have picked up the incident.

A. What is a SOC?

As mentioned previous, a SOC is a centralized hub for all network event monitoring and incident response related to cybersecurity events and incidents. SOCs are critical to all organisations when it comes to detecting, investigating and reporting on various malicious activities that occur. According to [2] there are five reasons why an organisation would want to introduce a SOC that is for proactive detection; threat awareness; vulnerability management; awareness of hardware and software assets; and log management.

Proactive detection is having the ability to see when a system's hardware or software has been breached by an unauthorized person. The unauthorized person could be internal (employee) or external.

Threat awareness is having the ability to find vulnerabilities in hardware and software that are used within an organisation. Assigning a score to each vulnerability allows the organisation to know which vulnerabilities are critical and which vulnerabilities should be addressed first. The vulnerability scoring should be assigned using the Common Vulnerability Scoring System (CVSS) scoring mechanisms as found on [3].

Vulnerability management is having the ability to handle the vulnerabilities that are found. There are four ways of managing vulnerabilities that are risks [4] namely risk

acceptance, avoidance, limitation and transfer. Risk acceptance is applied in cases where avoidance outweighs the risk itself and in this case an organisation may choose to accept that the particular hardware or software has a risk that can be exploited by attackers. Risk avoidance is applied in cases where an organisation is able to stop using the particular hardware or software that carries the risk. Risk limitation is applied in cases where an organisation can limit its dependency on a particular hardware or software that carries the risk. Risk transfer is applied in cases where the organisation can transfer the consequences of the risk to another organisation.

Awareness of hardware and software assets is having the ability to know which assets are critical and which ones are not so critical. Also knowing the full landscape of the IT environment.

Log management is having the ability to centrally manage all logs from the IT environment and being able to views trends and statistics coming from the logs.

II. SOC TOOLS

In order for a SOC to function, it needs tools. This sections describes some of the type of tools that should be found inside a SOC and what they can be used for.

A. Security Information and Event Management systems

A Security Information and Event Management (SIEM) is a tool that statistically looks at the events from different network sources such as hosts, servers and network endpoints in order to have a common approach to event modelling and security evaluations based on security metrics thereby providing risk analysis procedures [5]. A SIEM tool collects, analyses and reports on log data. A SIEM tool has effective graph generation techniques taking into account known attacks, new attacks, and even zero-day attacks. A SIEM tool will make use of analytical modelling and interactive decision support to provide the relevant security solutions.

B. Threat Intelligence Tools

With a SOC there are different types of tools that can be used to gather threat intelligence. The following are some threat intelligence tools that can be used.

The first is Live News. Live News allows the SOC personnel to view the latest news that has to do with cybersecurity breaches. The latest news might not affect the organisation or its partners to whom the SOC belongs but the information broadcasted can be used as threat intelligence in the event that the victim has a similar business model or similar IT technologies.

The second is Social Media, because hardware and software vendors communicate new vulnerabilities discovered in their products through social media platforms. There are Threat Intelligence platforms that can visualise real-time threat intelligence from social media feeds [6]. A SOC must have access to real time visualization of social media feeds from vendors of their hardware and software products.

The third is a threat Intelligence Platform. A Threat Intelligence Platform is a centralized tool that can pulls feeds from different threat databases based on an organisation's software and hardware assets [7]. This tool must also be able to use the CVSS scoring and Common Vulnerabilities and Exposures (CVE) to classify potential threats.

Lastly, information that is created as output from a SOC. That is data that is generated from the SOC such as vulnerability assessments and reports from investigation or SIEM graphs can be used as threat intelligence to have an idea of vulnerabilities and risks or the type of cyber-attacks that exist in the organisation's IT environment. All tools used in the SOC generate information and that information can be re-used as threat intelligence.

C. Vulnerability Assessment, Investigative & Forensic Tools

The following are types of tools that can be used for investigation and forensics.

Port scanning is a type of tool that can be used to determine the open and closed ports on hosts, servers and network devices. Unused Transmission Control Protocol over Internet Protocol (TCP/IP) ports should be closed. This can be seen a host or server hardening.

Wi-Fi vulnerability assessment is a type of tool that can be used to determine the weaknesses and vulnerabilities that are found in the Wi-Fi network such as testing the encryption algorithm and authentication methods.

Traffic capture and analyser is a type of tool that can be used to capture network traffic.

Sandbox is an IT environment that is similar to the organisation's IT Environment where malware can be executed in isolation without damaging the organisation's actual IT systems. A virtualized network or hypervisor environment can be used as a sandbox tool. Sandboxing allows the SOC to understand what the malware can do if executed in the organisation's IT environment.

Vulnerability Assessment tool is a type of tool that assess websites, operating systems (hosts, servers and network end points) and applications for known vulnerabilities. For applications and operating systems the vulnerability is mainly calculated using the CVSS version 3.0 scoring namely the base, temporal and environmental metrics. For website vulnerabilities the OWASP (Open Web Application Security Project) publishes a report every three years that details the top 10 most widely exploited web application vulnerabilities during that period [8].

Social Engineering assessment tool is a type of tool that can be used to assess the level of cybersecurity awareness within the organisation. There are specific tools that can be used to send emails with attachments or links to employees to test employees' response to suspicious emails.

D. SOC Tool Storage

All SOC tools that are used for investigation or reverse engineering should be store in a central location that can only be access by SOC personnel. A SOC may be mature enough to develop its own tools. These tools should also be stored in the same location. According to [9], the United States' Central Intelligence Agency was hacked and close to 9 000 documents and cybersecurity tools were published on Wikileaks. Therefore it is important that the storage area where SOC tools are kept should be encrypted. And the

machine or virtual machine that houses the tools should be periodically assessed for vulnerabilities. This machine should not be accessible from outside the SOC and preferable it should be on an isolated network.

III. SIEM Data Pulling/Polling Architecture

Technology integration can be direct or indirectly integrated. Direct integration means that technologies can communicate without requiring a third party technology. Indirect integration means that the two technologies will required a third party technology to communicate. Usually newer technologies and technologies nearing their End-of-Life (EoL) can be directly integrated while legacy technologies can only be indirectly integrated.

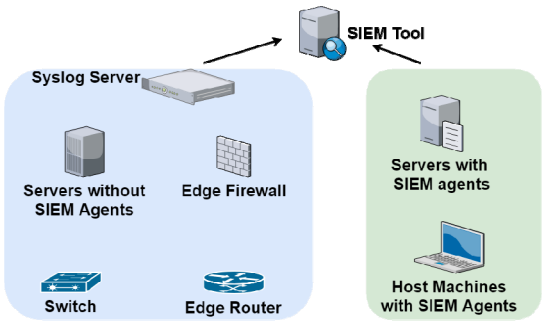


Figure 1: SIEM Agents Polling or Pulling Architecture

Fig.1 shows that in order to pull or poll incidents and events from a server and/or user host machine SIEM agents will have to be installed on the servers and user host machines. Servers that are running legacy applications will more likely have outdate operating systems on which the SIEM agents are not supported. Organisations usually do not have user host machines that are running outdate operating systems, if that is the case the organisation should consider upgrading these host machines or running them in secure virtual containers. All devices that are connected on the network that can install the SIEM agent on them will have send data directly to the SIEM tool.

All other network devices that cannot have the SIEM agent installed will send data to the SIEM indirectly. A network end point can be classified as any device that is on the network that is not a server or user host machine. Such as device could be a Switch, Firewall, Router, Intrusion Detection System (IDS) or Intrusion Prevention System (IPS) and Switch Port Analyser (SPAN). These devices cannot have agents installed on them in most cases, there they will need to forward syslog entries on port 514 to a central syslog server.

1) *Syslog Server*

On the syslog server the SIEM agent can be installed on it to pull/poll data into the SIEM tool. In the test case presented in this paper, the syslog server had several folders to store and categorize the type of syslog data received. The categories were namely, Firewall, Layer 2 Switch, Router (this includes Layer 3 Switches), Windows Servers, Linux Servers, Other Servers (this could include UNIX) and Network based IPS or IDS.

2) *Agent Installation*

The SIEM agents on the host machines and servers should be installed as a service on Windows and as a daemon on Linux machines. The Windows service or Linux daemon should be configured to always automatically run on boot-up.

IV. Types of Events to Monitor

Fig. 2 shows all the Windows Event logs. Not all events on a machine should be monitored. Also the events that are monitored on a Windows machine are different to those on a Linux or UNIX machine.

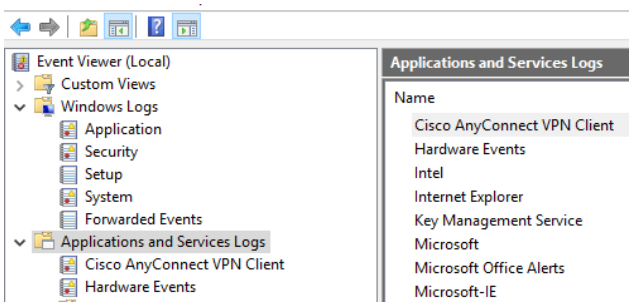


Figure 2: Windows Event Logs

A. *Windows Events to Monitor*

In this case study the following events were monitored on Windows machines. On the domain controller it is important to monitor the Windows Security logs. On application servers it is important to monitor the Windows Application Logs which gives information on the installation and uninstallation of applications. Applications and Services Logs allows the SIEM tool to capture data from application specific servers such as Domain Name Server (DNS), Document Sharing, Databases, Anti-Virus, etc. can be monitored.

B. *Linux Events to Monitor*

Fig. 3 shows all the Linux Event logs. On Linux, the operating system and applications store the log files in the /var/log/ location. In order to preserve the integrity of the logs and for security reasons the permission of the /var/log/ folder must be reserved to privilege users. That is only privilege users can write to the folders, subfolders and files. The "syslogd" and "rsyslogd" are the dedicated logging processes. The two main primary logging systems are /var/log/messages and /var/log/syslog.

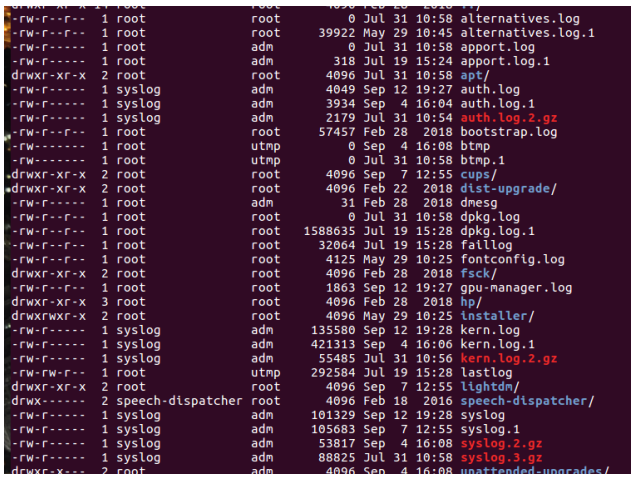


Figure 3: Linux Event Logs

In this case study the following events were monitored and the type of server is also given:

- /var/log/messages: General message and system related logs
- /var/log/auth.log: Authentication logs
- /var/log/kern.log: Kernel logs
- /var/log/cron.log: Crond logs (cron job)
- /var/log/boot.log: System boot logs
- /var/log/secure or /var/log/auth.log: Authentication logs
- /var/log/<application>: Logs for different applications depending on the application name.

C. Network Devices Events to Monitor

At a network level the events that will be monitored depends on the type of the device. Table 1 shows the type of events that were monitored per network device type.

TABLE 1 EVENTS MONITORED FROM A DEVICE TYPE

Device Type	What to Monitor
Firewalls	Session data with its IP 5 – tuples. The 5 tuple being source IP, destination IP, source port, destination port and transport layer protocol. Transaction data that is extracted documents, images, etc should be logged.
Intrusions Detection Systems /Intrusions Prevention Systems	Alert data that is when rules or anomalies are flagged.
Web Filters	Session data with its IP 5 – tuples. Transaction data that is extracted documents, images, etc should be logged.
Email Filters	Session data that is connection events. Statistical data should also be collected.
Endpoint Security (Antivirus & Antimalware)	Extracted data, that is the actual malware that was quarantined.
Virtual Private Network Concentrators	Authentication, authorization and auditing should be logged. Also the session data with its 5 – tuple.
Switches & Internal Routers	Administrative access to the devices, performance logs and type of traffic flowing through should be logged.
Edge Routers	Session data with its IP 5 – tuple.
Domain Name Servers	Transaction data that is queries or responses. Request for external domain names that are known to carry malware.
Authentication, Authorization & Auditing Servers	Alert data that is successful and failed authentication and authorization events. Tracking a successful event at the end of multiple failed events.
Dynamic Host Configuration Protocol Server	Transaction data that is IP assignments.

V. INTEGRATION POINTS

A SOC has people, processes and technologies. These must align with the parent organisation's people, processes and technologies.

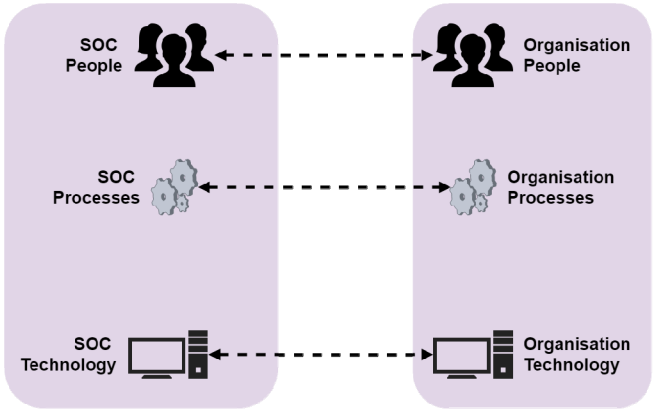


Figure 4: Integration of Poeople, Processes and Technology

A. People Integration

Fig. 4 shows that the people in the SOC must align and work with the people in the organisation. People skills are required to do this. According to [10] graduates engineers need to know that 64 percent of their salaries pays for their communication skills, not their engineering competencies. Communication is one of the key factors when integrating people from different roles in an organisation to work together.

The main challenge here is that cybersecurity is being introduced in an organisation and as a result the key stakeholders of system technologies and processes do not understand their roles with regards to cybersecurity. In the authors' experience the introduction of SOC people into an organisation and the attempt to integrate the SOC personnel and the organisation's key stakeholders was perceived at face value to the key stakeholders as a way in which they will be monitored closely. That is the key stakeholders of system technologies and processes, viewed the SOC personnel as people who will be watching their every move. In developing economies such as South Africa, job security is an important skill set [11]. When a new section is introduced the default view is they are going to take our jobs away and for fulfilment contractors or partners they view the new introduction as the people who will make our contract to be no longer needed. It is therefore important to have kick-off meetings with all the relevant system technologies and process stakeholders in order for them to understand that SOC personal are there to make their security job a lot easier. In particular the SOC manager must be able to clearly communicate the roles and responsibilities of the SOC and its cybersecurity related activities in an understandable language [12]. The SOC manager must be able to make the system technologies stakeholders understand that the SOC is there to ensure the organisation's IT systems, business data, employee data and customer data is secure from cyber criminals. The soft skills of the SOC manager are important for further communication and integration. The system technology owners can guide the SOC to which systems they know houses critical information required to ensure the business continues to function.

B. Technology Integration

A SOC has its own technology. Usually this is the latest versions of software or hardware. On the other hand an organisation has wide range of technology. An organisation's technology can be the latest, approaching EoL or legacy. An

organisation may accept the risk of running legacy technology for reasons such as the cost of acquiring newer technologies does not represent the Return on Investment (RoI), the technology is no longer manufacturer, etc. All these technologies have to be integrated with the SOC technologies.

C. *Process Integration*

A process in a SOC would be the steps that are followed in order to resolve an event. While a process in an organisation would be the steps taken to resolve a help desk call given the call's priority. In order to provide reporting and understand the RoI of a SOC, the SOC processes and the organisation processes have to be integrated. That is SOC processes have to conform to the similar steps as the organisation and as far as possible and be able to use the same systems.

VI. Processes and Procedures Found in an Organisation

A. *Change Advisory Board*

A Change Advisory Board (CAB) is part of the IT Service Management (ITSM) that upholds the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) 20000 standard [13]. A CAB is a group of key stakeholders such as subject matter expects and key managerial personnel who are responsible for evaluating changes to the IT environment. Most organisations consider the main role of the CAB to be to authorize the change but it should rather be to provide advice for the change. Another important role of the CAB is to evaluate the risk and mitigation techniques for a particular change together with the potential threat to how a business can be disrupted.

Firstly, by the very definition of the CAB and its responsibility within an organisation, it is important that a member of the SOC be part of CAB. Secondly during implementation and integration of the SOC technologies, it is important that the SOC technologies integration with the organisation's IT environment go through the CAB.

B. *Emergency Change Advisory Board*

An Emergency Change Advisory Board (ECAB) has the same roles as a CAB however this board is a subset of the members from the CAB and is only schedules to address urgent matters so that no unacceptable delays occur [13]. It is important that a member of the SOC team be part of the ECAB as well. Also some SOC security recommendations from investigations may require emergency changes to the IT environment these changes should be processed by the ECAB in order to mitigate potential vulnerabilities or prevent further breaches in an organisation's IT environment.

C. *Incident Management*

Fig. 5 shows that there are 4 methods by which an event can be logged as an incident [14]. Employees of the organisation and cybersecurity expects with the SOC can log calls for the incident to be investigated.

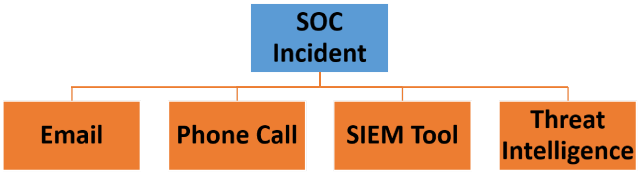


Figure 5: Tools used to Log a Cybersecurity Event

Another integration point between the SOC and the organisation is the introduction of a Cyber Security Team Support. This team is responsible for handling cybersecurity incidents and events logged against the SOC for investigation and resolving [14]. Fig. 6 shows the cybersecurity incident life cycle from detection through to reporting and feedback within the organisation.

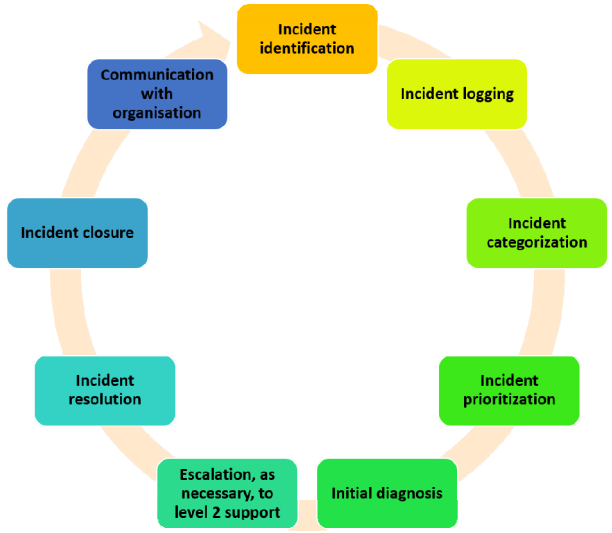


Figure 6: Incident Management Lifecycle

According to [14] there are 6 stages of an incident as shown in Fig. 7.

Figure 7: Six Stages of an Incident

When an incident is first received it is considered new, that is it has not been assigned to the Cyber Security Team Support. Once the incident is assigned it is then in progress in terms of investigation. If there is not enough information or other incidents have a higher priority, an incident can enter the holding or pending stage. An incident will then enter the resolved and finally the closed stage. From the closed state the playbook is updated and recorded in the SOC report.

Every incident should have the correct severity level assigned to it. Table 2 shows the priority matrix for assigning the severity to cybersecurity incidents.

TABLE 2 INCIDENT PRIORITY MATRIX

Incident Severity	Description
Low-priority	No disruption to users or the business and as a work around can be used.
Medium-priority	A few staff or systems are affected.
High-priority	A large number of staff or systems are affected. Potential financial impact on the organisation.

D. SOC Playbook

An important resource in the SOC for incident investigation is a SOC playbook. A playbook contains sub-processes that are step by step guides using distinct systems in order solve an incident [15]. Furthermore a playbook is an unrestrictive collection of previously seen plays (reports and methods) used to detect and respond to security incidents. A playbook is a known error database. The SOC Playbook is a living document which should be updated from time to time.

While investigating a suspicious event the playbook and checklists should prepare the analyst to properly investigate a security event and address the all-important who, what, when, where, why, and how questions that comprise a suspicious investigation.

VII. Reporting on Events and Incidents

Depending on the organisation a SOC report should be provided periodically depending on the need and requirement. The SOC report should provide a very complete view of the statistics of the events and incidents investigated and seen by the personnel in the SOC. The SOC report should also provide the location and identity of critical Windows/Linux host systems. The report should contain Security Incident Investigation Procedures.

VIII. Conclusion

Building a SOC for an organisation is an expensive exercise. However most organisation do not consider the challenges that comes with integrating a SOC into the organisations, in terms of the people, processes and technologies.

As demonstrated in the paper, there are integration challenges that comes with incorporating a SOC into the organisations. As far as possible this paper provided ways in which the challenges can be mitigated or addressed. In order to fully realize the potential of a SOC and allow the SOC to fight against cybersecurity attacks the integration of the SOC's personnel, technologies and processes needs to be correctly integrated with the organisation's personnel, technologies and processes.

In future it would be great to measure the integration levels between the SOC and its organisation. It is not possible to have a 100% integration therefore it is important that the next stage of the research would be to provide information on the minimum level of integration required to allow the SOC to begin functioning to its full ability.

References

[1] Verizon, "2018 Data Breach Investigations Report 11th edition," March 2018. [Online]. Available: http://www.documentwereld.nl/files/2018/Verizon-DBIR_2018-Main_report.pdf. [Accessed 13 September 2018].

[2] J. Wilkinson, "5 Reasons You Need A Security Operations Center (SOC)," 15 May 2018. [Online]. Available: https://www.lewan.com/blog/5-reasons-you-need-a-security-operations-center-soc.

[3] FIRST, "Common Vulnerability Scoring System v3.0," June 2015. [Online]. Available: https://www.first.org/cvss/cvss-v30-user_guide_v1.4.pdf.

[4] B. Peterson, "4 Ways to Manage Risk on Web Projects," 13 March 2017. [Online]. Available: https://gracefulresources.com/4-ways-to-manage-risk/. [Accessed 13 September 2018].

[5] C. Andrey and K. Igor, "Attack modeling and security evaluation in SIEM systems," International Transactions on Systems Science and Applications, vol. 8, pp. 129--147, 2012.

[6] J. Mtsweni, N. A. Shozi , K. Matenche, M. Mutemwa, N. Mkhonto and J. J. Jansen van Vuuren, "Development of a semantic-enabled cybersecurity threat intelligence sharing model," in Conference on Information Communication Technology and Society (ICTAS), Boston, 2017.

[7] J. Mtsweni, M. Mutemwa and N. Mkhonto, "Developing a cyber threat intelligence sharing platform for South African organisations," Journal of Information Warfare, vol. 15, no. 3, pp. 56-68, 2016.

[8] A. van der Stock, N. Smithline and T. Gigler, "OWASP Top 10 - 2017 The Ten Most Critical Web Application Security Risks," 20 October 2017. [Online]. Available: https://www.owasp.org/images/7/72/OWASP_Top_10-2017_%28en%29.pdf.pdf.

[9] S. Khandelwal, "WikiLeaks Exposed CIA's Hacking Tools And Capabilities Details," 7 March 2017. [Online]. Available: https://thehackernews.com/2017/03/wikileaks-cia-hacking-tool.html.

[10] S. Pneena and R. J. Carol, "A message from recent engineering graduates in the workplace: Results of a survey on technical communication skills," Journal of Engineering Education, vol. 90, no. 4, pp. 685--693, 2001.

[11] F. A. Philip, "Job Security in Developing Countries: A Comparative Perspective," Ife Journal of International and Comparative Law, pp. 51-75, 2016.

[12] D. Brecht, "The Skills and Experience Needed to Support A CSIRT, SOC or SIEM Team," InfoSec Resources, 7 February 2018. [Online]. Available: https://resources.infosecinstitute.com/skills-experience-needed-support-csirt-soc-siem-team/#gref. [Accessed 13 September 2018].

[13] S. Watts, "ITIL Change Management & the CAB," BMC, 28 December 2017. [Online]. Available: https://www.bmc.com/blogs/itil-change-advisory-board-cab/. [Accessed 13 September 2018].

[14] BMC, "ITIL® Incident Management," 22 Deccember 2016. [Online]. Available: https://www.bmc.com/guides/itil-incident-management.html.

[15] G. Sanker, "What's an ITIL CAB? A Simple Explanation," itsmtransition, June 2013. [Online]. Available: https://www.ucisa.ac.uk/-/media/files/members/activities/itil/servicetransition/chanage_management/itil_a%20guide%20to%20cab%20meetings%20pdf.ashx?la=en. [Accessed 13 September 2018].

[16] F. Bharathram, "Open Source Playbooks," IncidentResponse, 26 March 2018. [Online]. Available: https://www.incidentresponse.com/open-source-playbooks/. [Accessed 18 September 2018].

Renewable Energy Sources and Grid Integration in Ghana: Issues, Challenges and Solutions

K. Akom
Elect./Electronic Eng. Dept
University of Johannesburg
Johannesburg, South Africa
kingsleyakom@yahoo.com

M. K. Joseph
IEEE Computer Society
South Africa Chapter Chair
Johannesburg, South Africa
josephmeera@gmail.com

T. Shongwe
Elect./Electronic Eng. Dept.
University of Johannesburg,
Johannesburg, South Africa
tshongwe@uj.ac.za

***Abstract*—Due to the high cost of fuel and as unfavourable weather conditions which have affected power generation in Ghana, the country has experienced power crises for the past seven years. Renewable Energy Resources (RES) like wind and solar are being considered by many countries as alternatives for the energy requirements of the country. Ghana's Energy Commission's (EC) report in October, 2017 indicated that, RE contributes 2MW of power to the country's energy mix, representing about 1.73% of the total installed capacity. However, the current EC's energy policy has projected 600MW power through RE in 2030. 340 MW from solar and 260 MW from wind energies. The then Ministry of Energy through the Energy Commission started the rooftop PV programme implementation in early 2016 in some government institutions. The main aim of the rooftop programme was to produce about 200 MW maximum load respite on the national grid as a medium term programme through PV solar technology. However, RE and grid integration has various issues and challenges, large scale RE power generation are mainly connected to the transmission systems and small-scale generation are mostly linked with the distribution system. Direct integration these systems poses a lot of challenges and issues. This paper examines the main issues and proposes some probable solutions for future RE generations and integration.**

Keywords—power electronic technology, renewable energy source, grid integration, generation, transmission, distributed

I. INTRODUCTION

Usage of RE has gradually become an alternative source of power all over the world, most especially in developed and developing countries due to the fast exhaustion of fossil fuels [1]. The high cost of fuel for various electrical power plants, bad weather conditions and the related environmental challenges. Owing to the intermittent nature of RE sources, its integration into the main grid always leads to issues in both non-technical and technical challenges. To increase the renewable energy contribution in Ghana, the government as at March, 2017 had issued provisional wholesale power supply licenses to some 90 prospective IPPs to produce 5,000 MW of electricity from the RES available in Ghana [2, 3]. 60 out of the licenses issued proposed solar photovoltaic (PV) generation with 3,000 MW capacity of energy supply. 25 of the companies were licensed to produce solar energy with 3 companies issued with solar PV construction project license. 4 companies were issued with Construction Permit on wind energy to produced 225MW power [4].

In an effort to contribute to the national goal of universal access by 2020 and promote socioeconomic development through Productive Use of Energy (PUE), the second phase of the National Rooftop Solar Programme which focuses on providing electricity to rural households in off-grid communities which might not have access to electricity by 2020 and beyond was piloted in 2017. Solar PV systems of 500Wp capacity each were piloted in 200 rural households in 16 off-grid communities in the Abetifi and Mpraeso constituencies of the Eastern Region. Partial funding was secured from the Skills Development Fund (SDF) of COTVET 91 to train 250 technicians for designing, installation and maintenance of the solar PV system. However, ECG put the scheme on hold in the first quarter of 2017 for alleged fear of reduction in their revenue generation. The Energy Commission in collaboration with ECG and PURC are therefore currently working on resolving this challenge [2, 4].

The adoption of solar and wind energy systems has become a major challenge for most utility providers due to the potential negative influence it will pose on the power network operations system [2]. Another challenge on solar and wind energies is the fluctuation and intermittent power supply which has negative impact on the operation of the electric network, feeders [3] and nodes in the power network system [4]. Additionally, the variations and intermittent nature of solar and wind energies do not permit for using them in the process of planning and generation of electricity. The paper will examine all these challenges and provide probable solutions to those issues and challenges.

A. Objective

The main objective is to investigate the Renewable Energy Sources and Grid Integration in Ghana, explore the key issues and suggest some possible solutions.

B. Methodology

The methodology used for the studies are both qualitative and quantitative approach. In all, 42 questionnaires were sampled and analysed. Managers of Ghana's electrical energy systems were contacted and interviewed. They include engineers, technician engineers, technicians and artisans in Electricity Company of Ghana, Ghana Grid Company, Northern Power Distribution Company and some independent power producers.

II. DATA COLLECTED AND ANALYSIS

Two main tools were adopted in this regard, the data was collected through survey questionnaires. Since that was easy and less expensive according to Seliliz [5] than personal interviews. The questionnaires were developed to capture basic and technical all-inclusive data about the current state of energy situation in Ghana. The questionnaires were delivered by hand to each respondent and collected by the researcher. A second questionnaires were designed in a form of semi-structured interviews for the managers of the above mentioned companies in both close and open questions. Each interview lasted for about 45 minutes [3, 5]. Validity and reliability were ensured in the designing process of the questionnaire [5]. Frequency tables and descriptive statistics were created to display results with respect to each of the research questions. The analysis were than through inspection and checks of the surveys for wholeness and exactness, open coding and inputting data into a database in SPSS, and final execution of the descriptive responses in accordance with the descriptive statistics and frequency distributions.

A. Development and Distribution of Instruments

The questionnaires and the structured interview were developed to collect data for the following.

- The Current State of energy Situation in Ghana
- Renewable Energy projects in Ghana
- Overview of Renewable Energy projects in Ghana.

In arriving at the questionnaires and structured interview used for this work, series of tried ones were developed and given to colleagues and they gave a lot of valuable contributions and all these were incorporated in the questionnaires and structured interview to make sure that they were eliciting the right answers. Structured interview is the type of interview which is planned and written by the researcher. It guides him to look for the appropriate questions required for the interview. The interview was granted to the managers of VRA, ECG, GridCo and IPPs. The researcher went round administering the instrument in order to avoid the incidence of slow return rate. This was not an easy task since it involved extensive traveling across the country within a short time.

B. Quantitative Analysis

Table 1 Current state of energy supply in Ghana

	Frequency	**Percent**
Sufficient	10	18.9
Insufficient	30	56.6
Highly insufficient	13	24.5
Total	**53**	**100.0**

Table 1 shows that the majority (n=30, 56.6%) of the respondents indicated the current state of energy supply in Ghana is insufficient. Additionally, 13 (24.5%) of the respondents think that the current state of Ghana's energy supply is highly insufficient. From the responses, it can be concluded that current level of energy supply in the country is insufficient.

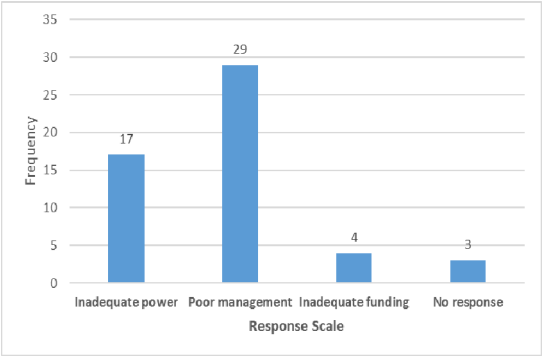


Figure 1 Causes of energy shortfall

Concerning the possible causes of the shortfalls in energy supply, the outcome as shown in figure 1 depicts that just about half (n=29) of the respondents attributed the causes to poor management whereas 17 respondents suggested inadequate power as the main causes of the shortfalls in energy supply in Ghana. The responses imply that mainly, poor management of the energy sector can be blamed for the current state of energy supply in the country.

Table 2 Degree of contribution from independent power producers

	Frequency	**Percent**
High	11	20.8
Moderate	7	13.2
Low	22	41.5
Very low	10	18.9
No response	3	5.7
Total	**53**	**100.0**

Enquiries about the current degree of contributions from independent power producers were made. From the table 2 it could be seen that 22 being 41.5% of the respondents indicated that there is low contribution to the energy supply by the independent power producers. However, 11 (20.8%) of the respondents suggested that they contribute highly to the country's energy supply needs. From the foregoing, it can be concluded that the contributions of the Independent Power Producers (IPPs) in Ghana is low.

Table 3 IPPs role in fixing prices in the energy sector

	Frequency	**Percent**
Very high	10	18.9
High	7	13.2
Moderate	4	7.5
Low	17	32.1
Very low	12	22.6
No response	3	5.7
Total	**53**	**100.0**

The results in table 4 concern the role of the IPPs in fixing prices in the energy sector. The responses suggest that the majority (n=17, 32%) of the respondents stated the producers have low involvement in the fixing of prices in the energy sector. That notwithstanding, 12 representing

about 23% of the respondents described it as having very low involvement in the pricing in the energy sector. The responses imply that the IPPs relatively do not play in any significant role in fixing prices in the energy sector. From table 4, descriptive statistics show that Solar obtained the highest mean score of 1.66 (SD=.478) among the other renewable energy resources. Followed up is Geothermal Energy with a mean score of (m=1.13, .342). Furthermore, respondents suggest Wind Energy (m=1.06, SD=.233) and Tidal Power (m=1.06, SD=.001). From the results, it can be concluded that solar energy remains the suitable renewable energy to augment the country's energy supply.

Table 4 Descriptive Statistics on renewable energy resources suitable for Ghana

	No.	Minimum	Maximum	Mean	Std. Deviation
Solar	53	1	2	1.66	.478
Wing Energy	53	1	2	1.06	.233
Geothermal Energy	53	1	2	1.13	.342
Biomass Energy	53	1	1	1.00	.000
Tidal Power	53	1	2	1.06	.233
Biofuel	53	1	1	1.00	.000
Valid N (listwise)	53				

The respondents were further requested to indicate which part of the country is suitable for the designated renewable energy development. Using the Kendall's coefficient of concordance for ranks (W^a) which estimates agreements between 3 or more respondents as they rank some subjects according to a particular characteristic as shown in table 6 it could be observed that the Northern part of Ghana received the highest mean score (m=4.39) hence ranked first as the most suitable place for the renewable energy development. Followed by the Western part (m=2.93) of the country.

Table 5 Regions

Region	Mean Rank
Northern Part	4.39
Southern Part	2.68
Central Part	2.50
Eastern Part	2.50
Western Part	2.93

Kendall's Wa=.508, χ2=83.268, df=4, Sig=.001

Reference to Kendall's (W^a) statistics in Table 5, the coefficients of concordance represents agreements, where 0 means no agreements at all, and 1 represents perfect agreement. A coefficient of 0.508 represents a moderate degree of agreement. Hence, can be concluded that there was significant agreement among the rankers (W^a=.508, $\chi 2$= (4)83.268, p<.05). Funding an important component required for the effective integration of RES into the grid system. Responses in Figure 2 regarding the funding approach required for the integration shows that respondents (n=31, 58.5%) wants the funding to be borne entirely by the government. However, 8 (15%) wants a partnership between the government as well as the independent power producers.

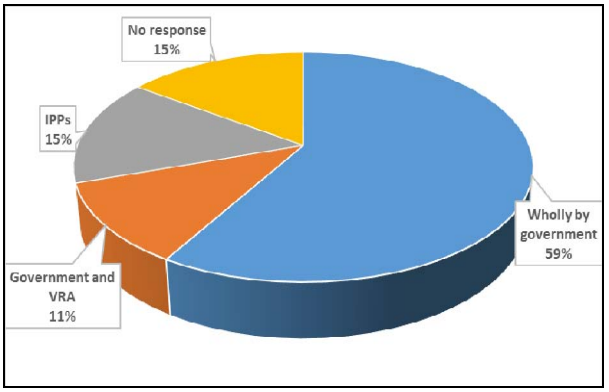


Figure 2 Funding approach required for integrating RES into grid system

From Table 6, responses on the potential of RES to alleviate many challenges facing the power sector suggests that with a mean score of 4.16 majority of the responses were towards the agreement end of the scale giving the impression that investing in the development of renewable energy sector has the inclination to improve the current energy situation in the country. Furthermore, the study showed that again, majority of the responses were around the agreement side of the scale (m=4.24) implying that respondents were in support of the assertion that educational institutions should be encouraged to generate their own power using the renewable energy sector. This means that getting government institutions particularly the educational institutions to develop their own renewable energy sources will lessen the burden on the national grid.

Table 6 Descriptive Statistics

	No.	Minimum	Maximum	Mean	Std. Deviation
RES has the potential to alleviate many challenges facing the power sector	45	3	5	4.16	.673
Educational institutions should be encouraged to generate their own power through RES	45	3	5	4.24	.712
Taxes on RES materials must be removed as a motivating factor	45	2	5	4.11	1.092
Rural Electrification should be solely RES	45	2	5	3.58	1.097

1=Strongly Disagree, 2=Disagree, 3=Not Sure, 4=Agree, 5=Strongly Agree

Additionally, majority of the responses (m=4.11, SD=1.09) gives the impression that the respondents agreed that taxes and duties on RES products should be removed to motivate institutions and individuals to invest in renewable energy sector. Also, regarding rural electrification, the trend of respondents (m=3.58, SD=1.097) suggest that most of the responses were geared towards the agreement end of the scale suggesting that respondents believe rural electrification should solely be on the renewable energy supply. This will help ease the strain on the national grid.

Table 7 IPPs are contributing immensely especially in the Rural areas

	Frequency	Percent
Extremely	11	20.8
Highly	14	26.4
Moderately	17	32.1
Quite/Reasonable	5	9.4
Somewhat	3	5.7
No response	3	5.7
Total	**53**	**100.0**

Assessment of the contribution of the IPPs to the current level of power supply in the country, the results in Table 7 shows that the majority (n=17, 32%) of the respondents view the contributions of the IPPs effort to the current supply of energy in Ghana to be moderate. Additionally, 14 (26.4%) of the respondent's think the IPPs are highly contributing immensely to power supply especially in the rural areas. The outcome of the results gives the impression that the independent power producers are not significantly contributing to the power supply in Ghana.

Table 8 IPPs have greatly influenced the Renewable Energy Sector

	Frequency	Percent
Not at all influential	7	13.2
Slightly influential	8	15.1
Moderately influential	25	47.2
Weak influence	10	18.9
No response	3	5.7
Total	**53**	**100.0**

The respondents were asked to assess the impact of the IPPs are having on the renewable energy sector. The results to that effect in Table 8 shows that the majority (n=25, 47.2%) of the respondents believe the IPPs are moderately influencing the renewable energy sector whereas about 10 (19%) of the respondents described the influence of IPPs in the sector as weak. From the responses it can be concluded that the IPPs are not significantly having much of an impact in the renewable energy sector as expected.

C. Qualitative Analysis

1. Number of Renewable Energy Projects in Ghana

The managers appeared divided regarding the specific number of renewable energy project currently underway in Ghana. However, majority interviewed were of the view that there are about between 4 - 20 renewable energy projects currently taking place in Ghana. Selected responses have been given below;

"We have four renewable energy projects in Ghana…"

"…I can't recollect off-head, but the last time I think it was 15 if my memory serves me right…"

"…There are about 20 separate companies currently in the country working on different renewable energy projects..."

The respondents were further asked to indicate where specifically the projects are located. The responses give the impression that the projects are spread across the country with major projects located specifically at Tema, Takoradi, Navrongo, Kpong and Bui. Some of the responses have been given below;

"Takoradi (Aboadze), Tema, Kpone, Akosombo (Ajena), Bui, Kpong (Akuse), Onyandze and Navrongo…"

"Takoradi, Tema and Accra"

2. Funding of the projects

Concerning sources of funds for the projects, the interview responses revealed the government is the main funding entity of the projects.

"The government is funding the projects with the taxpayer's money…"

3. Current production level of the renewable energy projects

The responses suggest that the perceived 20 renewable energy projects in operation in the country are contributing between 1.5 to 2.6 MW of power to the nation's energy pool.

4. Challenges associated with the integration into the National Grid

From the responses, the majority of the respondents maintained that the main challenge has to do with

difficulties in the mode of transmission whereas others also pointed to financial problems. Likewise, there were others who asserted that the risk is concerned with the increasing likelihood of flooding in coastal areas.

"It is financial problems that are troubling the projects…"

"…the mode of transmission into the national grid is now the major headache…"

5. Use of clouds for integration of renewable energy

Respondents were requested to indicate whether they consider the use of clouds as an alternative solution to the integration of renewable energy into the system. The majority of the respondents pointed out they are not considering the use of the resource as a means of integration into the system. Some of the responses to that effect have been outlined below;

"No! at the moment that is not what we are looking at…"

"To the best of my knowledge. NO! "No! not at all. No please"

III. ISSUES AND CHALLENGES OF RES GRID INTEGRATION

The main issues and challenges of RE and grid integration are classified in two forms. As a results of the interviews conducted in the study, technical and non-technical issues were emerged as the challenges [6].

A. Technical Issues

Some technical issues confronting the RES grid integration in Ghana are described as follows:

1. Poor Power Quality

The power quality is very poor in terms of harmonics, frequency and voltage fluctuations in various power stations in the country, especially Northern Electrification Distribution Company (NEDCo) Ltd

2. Transmission system

Less availability of transmission line to accommodate RES and the mode of RES transmission in Ghana is very challenging since they are mostly generated and transmitted in single phase.

3. Storage system

Due to the dynamic and complex nature of Power Storage and Grid Solutions, coupled with increasing renewable energy installations contributing power to an aging infrastructure in Ghana, energy storage has become an important focus [7, 8]. The study shows that energy storage in Ghana still remain a challenge in terms of frequency regulation and demand charge reduction or peak shaving.

4. Optimal placement of RES

Integration of large-scale distribution generation can be limited by these voltage variations in RES [9]. Limited amount of energy storage might not have the desired influence it is also not possible to install large amount of energy storage as this would increase the costs significantly. Therefore, it is required to optimally place and size energy storage.

B. NON-TECHNICAL ISSUES

1. Lack of technical skilled man power

Ghana lacks technical skilled man power in the energy sector. The sector also is saddled with huge debt which has affected the remuneration of its employers which attract the most suitable personnel. The government also has placed a ban on employment, so the managers are unable to employ qualify personnel for the sector [2, 3].

2. Poor management

Poor management also affects the operations of the energy sector. The sector is so politicized that, governments appoint the politicians instead of technocrats to manage the sector [3].

3. No Government support for the IPPs

Currently the IPPs lack resources to produce energy as expected of them and the government is not also position to assist in that direction. Successive governments have only shown interest in thermal power generation [3].

C. PROBABLE SOLUTIONS

In most developed countries all over the world solar and wind energy have become alternative energy sources [10] which Ghana must as well considering RES and distributed generators need new approaches for their operations in order to attain reliable and quality power supply [11]. The following propositions have been drawn as probable remedies for the integration of RES and grid systems in Ghana.

1. Minimized power fluctuations and intermittence in RES

The major challenge facing RES is power fluctuation and intermittence in power generation. This can be minimized to the barest minimum by distributing the RE in small units to large topographical areas by focusing on one area at a time [9]. The large solar PV system output power with high rating can be converted within ten minutes to avoid fluctuation and intermittence by natural weather conditions. Fluctuating and intermittence again can be reduced by using them for domestic activities [12].

2. Application of Power Electronic Technology

Power Electronic Technology (PET) is one of the contemporary technologies for distributing energy generated for the renewables and integrate them into the national grid system [13]. This mechanism is extensively applied globally as more efficient way of integrating RE into national grid system and ensuring reliable power with avoidance of fluctuation and intermittence. PET only requires semiconductor changers that have capacity for controlling higher voltages and for switching very fast without delay.

3. Appropriate Storage System is needed

Power output in large solar PV plants continuously fluctuate during the day time and this raises security concerns as fluctuating power results to unstable grid [14]. To resolve grid instability due to power fluctuation, solar PV plant owners will have to install storage systems, this leads to additional cost to installation or operational cost. [15].

4. The use of RES

RES are best utilized when the power is generated and use at the same time. For irrigation purposes, it is always advisable to feed the load at night or off-peak period through conventional grid.

IV. CONCLUSION

This paper investigated the specific challenges associated with grid integration of RES and appropriate recommendation are made based on observed data results after subjecting the data to a rigorous data analysis. The Northern part of Ghana received the highest mean score (m=4.39) hence ranked first as the most suitable place for the RE development in Ghana. To abate the electric power instabilities and its sporadic problems, power electronics devices were advised as the workable options. From the responses it can be concluded that the IPPs are not significantly having much of an impact in the renewable energy sector as expected. However, solar energy remains the suitable renewable energy to augment the country's energy supply. The application of dump load and proper storage of RE can also reduce drastically fluctuations and intermittence of solar PV systems [16, 17]. RE and grid integration in Ghana is possible and this can help the country to improve on its energy situation.

REFERENCES

[1] T. Ito, H. Miyata, M. Taniguchi, T. Aihara, N. Uchiyama, and H. Konishi, "Harmonic current reduction control for grid-connected PV generation systems", *International Power Electronics Conference*, Brighton, UK. pp 1695-1700, April 2010.

[2] A. S. Anees. "Grid integration of renewable energy sources: Challenges, issues and possible solutions", IEEE 5th India *International Conference on Power Electronics* (IICPE), August 2012.

[3] L. Adusei Energy security and the future of Ghana. Available *at: http://newsghana.com.gh/energy-security-and-the-future-of-ghana/* (2012) Accessed 3 July 2018.

[4] F.Meghdad, M. Asher, G. C. Klumpner and Y. Liangzhong, "Novel Integration of DFIG-Based Wind Generators Within Micro-grids", *IEEE Transactions on Energy Conversion, vol. 26, no. 3, pp 840-850,* September 2011.

[5] C, Seliliz, L.S. Wrightsman, and S. W. Cook. Research methods in social relations. *University of Michigan: Holt, Rinehart and Winston,* 1976.

[6] A. Douglas, Halamay, and T.K.A. Brekken, "Factorial Analysis for Modeling Large-Scale Grid Integration of Renewable Energy Sources", *PowerTech, 2011 IEEE Trondheim,* pp 1-6 June 19-23, 2011.

[7] S. Mukhtiar, K, Vinod, C. Ambrish, and V. K. Rajiv, "Grid Interconnection of Renewable Energy Sources at the Distribution Level With Power-Quality Improvement Features", *IEEE Transactions on Power Delivery*, vol. 26, no.l, pp 307-315, January 2011.

[8] S.K. Khadem, M. Basu, M.F. Conlon, "Power quality in grid connected Renewable energy system: Role of custom power devices", *International Conference on Renewable Energy and Power Qualities, Granada (Spain), pp 1-6, 23-25*, March 2010.

[9] G.C. Bakos, "Distributed power generation: *"A case study of small-scale PV power plant in Greece",* Applied Energy, vol. 86, pp 1757-1766, February 2009.

[10] A. Etxeberria, 1. Vechiu, A. Etxeberria, J.M. Vinassa, H. Camblong, "Hybrid Energy Storage Systems for Renewable Energy Sources Integration in Micro-grids: A Review", *International Power Electronics Conference (IPEC), pp 532-537,* March 2013.

[11] V. Foster J. Steinbucks, Paying the price for unreliable power supplies: in-house generation of electricity by firms in Africa. AICD. *Available at:* http://www.infrastructureafrica.org/system/files/library/2011/05/WP02_Own_generation.pdf_2008. Accessed 20 June 2018.

[12] D. Schulz, "Improved grid integration of wind energy systems", *Bulletin of the Polish Academy of Sciences Technical Sciences, vol. 57, no. 4, pp 311-315,* 2009.

[13] A. Amiri M. Zibaei *(*Granger causality between energy use and economic growth in France using geostatistical methods. *Munich Personal RePEc Archive (MPRA) Paper 36357*. 2012

[14] V. C. Ganti, Bhim Singh, S. K. Aggarwal, and T. C. KandpaI, "DFIG-based wind power conversion with grid power leveling for reduced gusts," *IEEE Transactions on Sustainable Energy, vol. 3, no. I, pp 12-20,* January 2012.

[15] A Omran, Walid, M. Kazerani and M. M. A. Salama, "Investigation of Methods for Reduction of Power Fluctuations Generated From Large Grid-Connected Photovoltaic Systems", *IEEE Transactions on Energy Conversion, vol. 26, no. 1, pp 318- 327,* March 201 I.

[16] Tidjani, F. Souleyman, C. Ambrish "Integration of renewable energy sources and the utility grid with the Net Zero Energy Building in Republic of Chad", *IECON 2012 - 38th Annual Conference on IEEE Industrial Electronics Society,* May 2012.

[17] D. Stroe, A. Stan, R.Diosi, R. Teodorescu, S. J. Andreasen. "Short term energy storage for grid support in wind power applications", 2012 *13th International Conference on Optimization of Electrical and Electronic Equipment (OPTIM),* April 2012.

Create Awareness of Electronic Waste hazards on the Environment in the Developing Regions using 21st century Learning Environments

Ramadile Isaac Moletsane
Faculty of Applied and Computer Sciences
Vaal University of Technology
Vanderbijlpark, South Africa.
ramadilem@vut.ac.za

***Abstract*— The expansion of Information and Communication Technology has prompted take-up in advanced education. 21st century learning situations are continually advancing because of uncommon pace of technological developments. The growing use of these learning environments has resultant increase in the generation of electronic waste. The growing number of populace in the developing countries with disposable income has attracted investments in information and communication technology into their countries. Institutions of higher education are unlikely to be left behind. The pervasiveness of electronic gadgets has led to increase in the generation of electronic waste. Electronic waste is a global concern due to its effects on human health and the environment. the aim of this narrative review paper is to explore the potential of 21st century learning environments in creating awareness about electronic waste negative effects on the environment. Findings reveal that developing countries know little about the harmful results of electronic waste on the environment and wellbeing. Based on the findings, therefore this study suggests that 21st century learning environments are the best method to combat electronic waste illiteracy in the developing countries.**

Keywords—21st century learning environments, information and communication technology, electronic waste awareness, higher education institutions, environment, health, developing regions

I. INTRODUCTION

Right from all levels of schooling to the Higher Education Institutions (HEIs) levels Information and Communication Technology is omnipresent. This trend captured the rest of the world. The 21st century learning environments (21st_cLEs) represent a modern development in pedagogy through the incorporation of ICT. The use of 21st century learning tools to create and support 21st_cLEs cannot be ignored. ICT in supporting 21st_cLEs offer both opportunity and challenges. Today's world business solutions insist 21st century skills which survive on 21st_cLEs. The resultant use of 21st_cLEs has led to generation of electronic waste. In particular e-waste represents one of the pressing challenges faced by today's world.

Electronic waste grows at the rate roughly from 3% to 5% every year [1]. In 2016 and 2017 about 45.7 million metric tonnes (Mt) and 49.8Mt of e-waste respectively was reported to be generated around the world with estimates of 49.8Mt for 2018 [2]. E-waste is one of the most significant (harmful) challenges that humanity is facing, due to its growing volume and toxicity [3]. This study is motivated by perceived lack of e-waste awareness in the developing countries— in particular African regions. The paper is structured as follows: In Section II electronic waste definition is provided then followed by the definition of 21st century learning environments in Section III. The opportunities and challenges offered by 21st century learning environments are discussed under Section IV. E-waste in the developed countries and E-waste in the developing world are discussed in Section V and Section VI respectively. The study concludes in Section VII with conclusion.

II. ELECTRONIC WASTE DEFINITION

There is no universally agreed upon definition of e-waste. Each country has come up with its own definition of e-waste [3]. This inconsistency has led to many definitions contained in e-waste legislations, policies and guidelines [4]. According to Step initiative [4] e-waste " is a term used to cover items of all types of electrical and electronic equipment (EEE) and its parts that have been discarded by the owner as waste without the intention of re-use". Grant, Goldizen, Sly, Brune, Neira, van den Berg and Norman [5] define e-waste as any end of life "equipment which is dependent on electric current or electromagnetic fields in order to work properly." E-waste is the name given to all electronic and electrical equipment which are at the end of its life [6]. For this study, e-waste is defined as all electric and electronic equipment discarded or unwanted by the owner, regardless of working state or not, that contains both toxic and valuable materials. There are 10 categories of e-waste according to the European Union directives on WEEE. E-waste is often used interchangeably with WEEE, i.e. Waste from Electrical and Electronic Equipment [7]. The classification is given in ascending order and identified by the following labels: major household equipment, minor household equipment, IT and Telecommunications tools, Electrical and electronic tools with the exclusion of large scale stationary industrial tools, Toys, leisure and sports equipment, medical appliances with the exception of all implants and infected products, monitoring and control tools, Automatic dispensers [8].

III. DEFINING 21ST CENTURY LEARNING ENVIRONMENTS

The 21st_cLEs is a learning situation that utilizes innovation to offer a variety of learning encounters to students. It extends the learning outcome past the bricks and mortar classrooms. 21st_cLEs includes something other than the utilization of innovation, yet students utilizing innovation to accomplish objectives uniquely in contrast to what was conceivable previously. The term learning environment or condition propose a place and space— school, classroom or library. Normally 21st century learning happens in physical areas like these. These physical areas

are portrayed by computers, smartboards, local area networks or laptops for each student [9]. 21st_cLEs cases likewise incorporate mobile phones such as smartphones [10].

IV. Opportunities and Challenges of 21st Century Learning Environments in Higher Education Institutions

21st_cLEs either physical or online are characterised by ICT equipment such as computers, laptops and electronic whiteboards to mention a few. ICT have revolutionised education in many new ways [11]. In the midst of 21st_cLEs undisputable positive impact on education the tools that support these environments create e-waste when they become obsolete or unwanted irrespective of their functional state [11].

A. Opportunities of 21st century learning environments

In the midst of the race across the world to transform education systems into 21st_cLEs that does not necessarily mean that one has to leverage technology to do so. According to Keengwe and Gergina [12] "technology is not a substitute for good instruction". The 21st_cLEs makes it possible for students to acquire skill demanded by the new knowledge economy [13]. Absence of these skills further widens the digital divide gap [14]. Eventually nations regress to be colonized in the web of digital divide by those who possess these skills. Table 1 shows some of the current and new innovation brought by the technology in our education systems [15].

Table 1 Current and Emerging modern 21st learning environments [15]

First order innovations	Second order innovations
Blogs and wikis	Augmented reality
Social networking sites	Simulations
Virtual learning environments	Digital games
Personal computers, tablets, smartphones	Console games
Interactive whiteboards	Remote response systems
Web applications	Mobile computing

Table 1 first order innovations refer to common 21st_cLEs while a second order innovation refers to the emerging 21st_cLEs. Many of first order innovations belong to Web 2.0 technology a term used represent internet tools that allows transfer of 21st century skills such as communication and collaboration [15]. According to Boholano [11] 21st_cLEs allow collaboration between teams or individual students everywhere around the world who never meet face to face. Videoconferencing as 21st_cLE support collaboration [16]. 21st_cLEs support a combination of some in-person learning and some online learning that can be described as hybrid or blended learning [17]. 21st_cLEs knows no space and time hence teaching and learning is available 24 hours and 7 days. Students can learn by engaging in on-site classes, online off-site classes or a mix of both modes. According to Keppell, Souter and Riddle [14] 21st_cLEs addresses the limitations of traditional teaching and learning environments which tends to be physical and characterised by mortar and bricks buildings. In these environments students are passive receivers of information from the lecturer relying from a prescribed textbook most of the time.

Utilizing networked mobile equipment, people can both access and add to a developing information base, they can capture, alter and distribute sound, video and pictures from anyplace and whenever, and by including their voice, they can impact worldwide discussions [18].

The use of 21st_cLEs is not limited only to education sector but also visible elsewhere such as in creating awareness of social challenges—e.g. health related issues, xhinophobia issues, discrimination at work and violence against children and women. 21st_cLEs make it possible for community and individuals' voices to be heard [19]. The following examples show how 21st_cLEs championed awareness initiatives and became a success story.

Digital Storytelling as 21st_cLEs tool was effective to raise awareness through the study conducted by Cueva, Kuhnley, Revels, Schoenberg and Dignan [20]. The motivation behind the investigation was to learn viewpoints of group individuals subsequent to observe the digital story made by group well-being professionals. The study demonstrated that the members found the digital story as an adequate, emotionally engaging way to increase viewer's cancer awareness and to begin discussions. The discussions filled in as a trigger for reflection, insight, cancer prevention and hazard decrease exercises. Another health related success of 21st_cLEs in raising awareness was from a study conducted by Briant, Halter, Marchello, Escareño and Thompson [21].

21st_cLEs— in particular digital storytelling was found to be a powerful health promotion method. It emerged from the interviews that digital storied had a potential for disease prevention awareness and education. 21st_cLEs— in particular mobile devices with multimedia capabilities found to increase student's use of mobile gadgets and to develop environmental awareness. The study participants learned approaches to keep clean environment and increased their consciousness of environmental concerns [22].

The power of 21st_cLEs in creating awareness of social challenges in the 21st century cannot be overemphasised especially in the developing countries. With reference to Salagar, Kulkarni and Gondane [19] it is critical to employ 21st_cLEs to develop awareness in the developing countries due to social self-consciousness, social barriers, gender inequality and educational system. Indeed 21st_cLEs is a disruptive innovation and it has changed the way students and communities learn and socialise [23].

B. Challenges of 21st century learning environments

The challenges of 21st_cLEs range from e.g. implementation challenges, e-waste, ergonomics, and data theft from e-waste. In this study e-waste generated from unwanted 21st_cLEs tools will be dealt with. E-waste has recently become worldwide concern due to ever-increasing

volumes and poisonous metals that compose electronic equipment [24]. E-waste when not environmentally friendly it poison the ecosystem and result severe consequences to human and animals wellbeing including poisoning of food chain. Toxins from e-waste materials can bio-accumulate in agricultural lands and present in grazing livestock's foods [5].

An examination led in China uncovered that soil pollution from airborne accumulation is the fundamental wellspring of toxic metals tainting in rice [25]. Literature reported that poison form e-waste rare metals and fumes when released into the environment caused deformities such as brain underdevelopment in minors and skin related diseases [5, 26]. Indeed e-waste is bad for both environment and the health of humans and animals when is poorly managed.

However e-waste is offers economic opportunities [27]. Behind those evil black clouds of e-waste fumes and devastating effects on wellbeing and environment buried a wealth. E-waste in 2014 generated about US$54 billion in plastics and other rare metals [2]. The next discussion provides the condition of e-waste in both developed and developing countries.

V. Electronic Waste and the Developed Regions

In many developing countries the producer or manufacturer of electric and electronic products is responsible for the disposal of obsolete products. The sad news is that three-quarters of the products sold to Europe are produced in the developing countries—China and India. So much responsibility measures make it difficult for poor developing nations [28]. When coming to recycling the developed countries boosts state of the art recycling centres and set targets for recycling. For example Europe has evolved regulations with its member states with the aim of recycling at-least 85% of generated e-waste by the year 2019 [6].

A. Japan and the electronic waste problem

The world's leading manufacturer of electronic products has a limited number of e-waste specific regulations. In 1998 Japan passed Specific Home Appliances Recycling Law (SHARL). In 2001 SHARL was amended and has since then ensured the treatment of home appliances [29]. Another strategy adopted by developed countries in order to manage e-waste is extended producer responsibility (ERP). One of the features of EPR is "the shifting of responsibility (physically and/or economically; fully or partially) upstream toward the producer and away from municipalities"[30]. ERP is often termed producer take back [31]. Most ERP laws require the consumer to bear a portion of the costs mandated with transportation and collection. According to Garlapati [6] ERP encourages producers to seek ways that reduce energy and eliminate toxic materials in all stages of e-waste producing equipment. By so doing the producer or manufacturer of the electrical and electronic equipment gains corporate credibility and reputation.

B. European Union and electronic waste challenge

The European Union (EU) has progressive and comprehensive set of regulations and laws on e-waste. The policy on e-waste in EU dated back in the nineties. Here are some of the regulations: The Waste Shipment Regulation (WSR) passed and amended in 1993 and 1997 respectively. The objective of WSR is to ensure that none of hazardous waste gets shipped to Organisation for Economic Cooperation & Development (OECD) countries [29]. OECD founded in 1961, represents a 34 member sovereign states to stimulate economic development and social policies among its members [32]. "A club of most richly countries"[33]. In 2003 a directive was passed for changing the product design and increasing recycling efforts of e-waste generating equipment and Restriction of the use of certain hazardous substances (RoHS) such as mercury in locally produced or imported equipment that generates e-waste. RoHS addresses the beginning of electric and electronic equipment life cycling by attempting to reduce or eliminate toxic materials [29].

VI. Electronic Waste and the Developing Regions

Notwithstanding legislation and laws e-waste management is a plague in the developing countries. The developing countries challenges on e-waste management are primarily due to: absence of policy and if any nonattendance to policy implementation processes [34].

Developing countries e-waste growth rates are higher than developed countries due to accelerated information and communication technology growth rates. Provide examples. There is a paucity of studies relative to developed countries in the developing countries [35]. According to Heeks, Subramanian and Jones [36] developing countries suffer from knowledge gap of e-waste. They further divided knowledge gap into three types: absence of statistics on the e-waste, lack of methods deemed to manage e-waste properly and the role of organisations that produces e-waste generating products in relation to e-waste reduction. The following discussions look into e-waste issues in some of the developing countries.

A. China and the electronic waste challenge

China regulation on e-waste is not working and most of the time not enforced. In China e-waste is poorly coordinated. There are about ten departments responsible for e-waste legislation related issues. The matters handled by these departments include imposing disposal fees, illegal imports of e-waste and making amendments etc. the major problem with these departments is lack of coordination. There is a limited crosstalk [28].

.

B. South Africa and the electronic waste problem

E-waste account 5% to 8% of municipal solid waste in South Africa and it grows at the rate three times faster than any form of waste in South Africa. the challenges of e-waste are the results of absence of public awareness and market based instruments, lack of recycling infrastructure relative to the developed countries, absence of e-waste collection systems , illegal disposal , transportation and sorting of e-waste. South African state departments— and owned enterprises— e.g.: Eskom, Transnet and South African airways independently manage their e-waste. Coordination could maximise the efforts to manage e-waste properly [37].

In South Africa most e-waste in large portions is secured workplaces, storerooms, homes et cetera [38]. As indicated by the examination directed by Wakuma (2014:1166) it was concluded that amassed e-waste in storerooms and homes is due to absence of e-waste awareness. South Africa produces 77000 tons of data and correspondence innovation electronic waste every year [39].

South Africa has some of the progressive laws on human rights as enshrined by the constitution of the country, but lack legislation specific to electronic waste. South Africa is a signatory of the Basel conversion. Basel tradition is an arrangement intended to decrease the development of unsafe waste and their transfer amongst nations and particularly to prevent exchange of this loss from the created nations to the creating nations [40]. South Africa isn't a signatory of Bamako tradition which infers that the nation can import electronic waste legitimately [41].

C. Nigeria and electronic waste challenge

In Nigeria like most developing countries there is an absence of specific regulation to e-waste. However there are some relegations dealing with general waste. The problem with these current laws on general waste is lax enforcement. There is a lack of formal facilities and methods to safely process e-waste. Given the absence of infrastructure to deal with e-waste most if not all e-waste is processed in the informal sector [34].

Problem with informal processing methods and site ranges from absence of workers' rights such as safety at work. Literature has shown that informal workers are exposed to harmful fumes at higher levels relative to formal workers. Heightened exposure to toxic materials eventually leads to chronic health problems and eventually premature deaths [27].

D. Ghana and the electronic waste problem

According to Daum, Stoler and Grant [42] Ghana e-waste activities generate about US$268 million per year and so enable about 200,000 people countrywide to put bread on their tables. While Ghana is a signatory of both Bamako and Basel conventions these international legislations are not in force. According to Daum, Stoler and Grant [42] the reason for Bamako and Basel conventions not in force is that they are both not yet incorporated to a law. In Ghana there is an array of environmentally related legislations but none specifically related to e-waste. Government has implemented some low scale awareness programs on e-waste and handling toxic materials. The success of these programs is limited by absence of environmentally friendly disposal methods. Most of e-waste is either stored or given to informal collectors who then sold it to the informal recyclers [42, 43].

VII. CONCLUSION

This paper discusses the challenges and opportunities provided by the 21st_cELs. It also discusses e-waste challenging issues in the developing countries. The findings reveal that e-waste is poorly managed in the developing countries. Absence of awareness on e-waste was also identified. The governments of the developing nations should launch awareness programs. 21st_cLEs is identified as an agent to relay knowledge and awareness specifically to e-waste workers. These tools have the potential to get information anytime and everywhere.

FUTURE RESEARCH AND LIMITATIONS

Africa in particular does not have much literature base on e-waste generated by electric and electronic equipment. There is a paucity of studies undertaken on e-waste and subsequently less is known and understood about e-waste. This study is limited to e-waste challenges faced by organisations —e.g.: higher education institutions, governments departments, etc. in the developing countries.

ACKNOWLEDGEMENTS

The author would like to thank all contributed to the success of this paper

REFERENCES

[1] A. Kumar, M. Holuszko, and D.R.C. Espinosa, "An overview on generation, collection, legislation and recycling practices," Res Conser and Recycling, pp. 32-42, 2017.

[2] A. Kumar, and M. Holuszko, "Electronic Waste and Existing Processing Routes: A Canadian Perspective," Resources, vol. 5, pp 1-19, 2016.

[3] Step initiative, "One Global Definition of E-waste," United Nations University, pp. 1-13, 2014.

[4] K. Grant, F.C, Goldizen, P.D. Sly, M.N. Brune, M. Neira, M. van den Berg, and R.E. Norman, "Health consequences of exposure to e-waste: a systematic review," Lancet Glob Health, vol. 1, pp 350-361, 2013.

[5] V.K. Garlapati, "E-waste in India and developed countries: Management, recycling, business and biotechnological initiatives," Renewable and Sustainable Energy Reviews, vol. 54, pp. 874-881, 2016.

[6] C. Frazzoli, c.e. Orisakwe, R. Dragone, and A. Mantovina, "Diagnostic health risk assessment of electronic waste on the general population in developing countries' scenarios," Environmental Impact Assessment Review, vol. 30, pp. 388-399, 2010.

[7] B. Xakalashe, "An overview of recycling of electronic waste PART 1," Norwegian University of Science and Technology and Mintek, Randburg, pp. 8-12, 2012.

[8] H. Montrieux, R. Vanderlinde, T. Schellens, and L. De Marez, "Teaching and Learning with Mobile Technology: A Qualitative Explorative Study aboutthe Introduction of Tablet Devices in Secondary Education," PLoS ONE, vol. 10, pp. 1-17, 2015.

[9] Pacific Policy Research Center, "21st Century Skills for Students and Teachers," Kamehameha Schools Research & Evaluation, pp. 1-25, 2010.

[10] H.B. Bo,"olano, "Smart Social Networking: 21 st Century Teaching and Learning Skills," Research in Pedagogy, vol. 7, pp. 21-29, 2017.

[11] J. Keengwe, and D. Gergina, "Supporting digital natives to learn effectively with technology tools" International Journal of Information and Communication Technology Education, vol. 9, pp. 51-59, 2013.

[12] A.L. Cloete, "Technology and education: Challenges and opportunities," Theological Studies, vol. 73, pp. 1-7, 2017.

[13] M. Keppell, K. Souter, and M. Riddle, "Re-Imagining Teaching for Technology-EnrichedLearning Spaces: An Academic Development Model," IGI Global, 2012.

[14] J. Groff, "Technology-Rich Innovative Learning Environments," pp. 1-30, 2013.

[15] H.A. Klenke, "The effects of interactive videoconferencing on elementary literacy," Graduate Research Papers, vol. 192, pp. 1-36, 2014.

[16] Y. Zhonggen, "Blended learning over decades," International Journal of Information and Communication Technology Education, vol. 11, pp. 1-19, 2015.

[17] J.A. Taylor, and D. Newton, "Beyond blended learning: A case study of institutional change at an Australian regional university," Internet and Higher Education, vol. 18, pp. 54-60, 2013.

[18] M. Salagar, P. Kulkarni, and S. Gondane, "Educating and creating social awareness for sensitive topics using mobile applications," Book Educating and creating social awareness for sensitive topics using mobile applications, pp. 335-339, 2013.

[19] M. Cueva, R. Kuhnley, L. Revels, N.E. Schoenberg, and M. Dignan, "DS: a tool for health promotion and cancer awareness in rural Alaskan communities," International Journal of Circumpolar Health, vol. 74, 2015.

[20] K.J. Briant, A. Halter, N. Marchello, M. Escareño, and B. Thompson, "The Power of Digital Storytelling as a Culturally Relevant Health Promotion Tool," Health Promotion Practice, vol. 17, pp. 793-801, 2016.

[21] H. Uzunboylu, N. Cavus, and E. Ercag, "Using mobile learning to increase environmental awareness," Computers and Education, vol. 59, pp. 381-389, 2008.

[22] Macarthur Foundation, "Digital media and learning," 2013.

[23] M.Q. Patton, "Qualitative research and evaluation methods," CA: Sage, 1990.

[24] J. Fu, Q. Zhou, J. Liu, W. Liu, T. Wang, Q. Zhang, and G. Jiang, "High levels of heavy metals in rice from a typical E-waste recycling area in southeast China and its potential risk to human health," Chemosphere, vol. 71, pp. 1269-1275, 2008.

[25] D.N. Perkins, M. Drisse, and T. Nxele, "E-Waste: A Global Hazard," Annals of Global Health, vol. 80, pp. 286-295, 2014.

[26] J.K. Park, L. Hoerning, S. Watry, T. Burgett, and S. Matthias, "Effects of Electronic Waste on Developing Countries," Advances in Recycling & Waste Management, vol. 2, pp. 1-6, 2017.

[27] Z. Wang, B. Zhang, and G. Dabo, "Take responsibility for electronic-waste disposal," Nature International Weekly Journal, vol. 536, pp. 23-25, 2016.

[28] U. Kumar, and D.N. Singh, "E-waste Management through Regulations," International Journal of Engineering Inventions, vol. 3, pp. 6-14, 2013.

[29] OECD, "Extended Producer responsibility, A Guide Manual for Governments," 2001.

[30] Clean Production Action, "How Producer Responsibility for Product Take-Back Can Promote Eco-Design," Book How Producer Responsibility for Product Take-Back Can Promote Eco-Design, pp. 1-6, 2007.

[31] C. Thevenot, "Inequality in OECD countries," Scandinavian journal of public health, vol. 45, pp. 9-16, 2017.

[32] C. Thevenot, "Iniquality in OECD countries," Scandinavian Journal of Public Health, vol. 45, pp. 9-16, 2017.

[33] H. Aborele, C, Urquhart, and S. Nudurupati, "E-Waste Policy in Developing Countries: The Context of Nigeria," Book E-Waste Policy in Developing Countries: The Context of Nigeria, pp. 1-12, 2015.

[34] S. Sthiannopkao, and M.H. Wong, "Handling e-waste in developed and developing countries: Initiatives, practices, and consequences," Science of the Total Environment, pp. 1147-1153, 2013.

[35] R. Heeks, L. Subramanian, and C. Jones, "Understanding e-Waste Management in Developing Countries: Strategies, Determinants, and Policy Implications in the Indian ICT Sector," Information Technology for Development, vol. 21, pp. 653– 667, 2015.

[36] R. Heeks, L. Subramanian, and C. Jones, "Understanding e-Waste Management in Developing Countries: Strategies, Determinants, and Policy Implications in the Indian ICT Sector" Information Technology for Development, vol. 21, pp. 653-667, 2015.

[37] ITWeb, and A. Moyo, "SA loses e-waste opportunity to foreign countries," Book SA loses e-waste opportunity to foreign countries, pp. 1-2, 2017.

[38] GreenCape, Q. Williams, K. Barners, S. Tawanda, and S. Smout, "Waste Economy," Book Waste Economy, pp. 10-57, 2017.

[39] F. Alias, M.B. Ishak, S.N. Zulkifli, and R.A. Jalil, "E-waste management: An emerging global crisis and the Malaysian scenario," Journal of Environmental Sciences, vol. 4, pp. 444-457, 2014.

[40] A.J.J. Mouton, and J.H. Wichers, "A Proposed Model for South Africa to Effectively Recycle," Book A Proposed Model for South Africa to Effectively Recycle, PP. 1-5, 2016.

[41] K. Daum, J. Stoler, and R.J. Grant, "Toward a More Sustainable Trajectory for E-Waste Policy: A Review of a Decade of E-Waste Research in Accra, Ghana," International Journal of Environmental Research and Public Health, vol. 14, pp. 1-18, 2017.

[42] M. Oteng-Ababio, "Electronic Waste Management in Ghana-Issues and Practices," Sustainable Development - Authoritative and Leading Edge Content for Environmental Management, PP 149-166, 2012.

Rhonda: the architecture of a multilingual speech-to-speech translation pipeline

Johannes A Louw
Human Language Technologies Research Group
Meraka Institute, CSIR
Pretoria, South Africa
jalouw@csir.co.za

Avashlin Moodley
Human Language Technologies Research Group
Meraka Institute, CSIR
Pretoria, South Africa
amoodley1@csir.co.za

***Abstract*—Speech-to-speech translation can be described as converting a speech signal from a source language into a speech signal of the same meaning or intent into a target language. This process is achieved by the coordinated cooperation of individual Human Language Technology components, where the most important components to a speech translation system are automatic speech recognition, machine translation and text-to-speech. In this paper we present and discuss the design and architectural building blocks of the Rhonda speech-to-speech translation system, as well as their interactions with each other to facilitate speech-to-speech translation in a reliable, scalable and possibly distributed manner.**

***Index Terms*—speech-to-speech translation, automatic speech recognition, machine translation, text-to-speech, multilingual**

I. Introduction

The goal of an automated *speech-to-speech* (S2S) translation system is to bridge a verbal language barrier that exists between speakers of differing first languages, and no mutual understanding of an auxiliary language. Verbal language barriers in communication are not only limited to multilingual societies and countries such as South Africa, but can manifest itself in monolingual societies and regions as well through for example: *migration*, *tourism* and *commerce* activities. The adverse effects of language barriers can have serious and wide reaching consequences on individuals and groups through various interactions, for example:

- *Healthcare*: patient understanding and satisfaction from primary and preventive care [1].
- *Law Enforcement and the Legal System*: ineffective communication causing delays and frustrations [2], [3].
- *Business*: the breakdown of communication in multinational corporations [4].

An automated S2S translation system can alleviate many of the negative consequences of language barriers in a cost effective and timely manner.

A speech-to-speech translation system is an interdisciplinary *Human Language Techology* (HLT), with the development of such a system requiring the skills and expert knowledge of teams of linguists, computer scientists and engineers. There are in general three main HLT components in an S2S translation system (other HLT components that may be incorporated into an S2S translation system in order to improve it's translation performance and naturalness): automatic speech recognition (ASR), machine translation (MT) and text-to-speech (TTS). These three technologies are combined in a manner that enables the S2S translation system to perform the speech-to-speech translation as follows:

- the ASR component recognises input speech and transcribes it into a textual form of the input *source language*,
- the MT component translates the transcribed text of the *source language* into text of the *target language*, and
- the TTS component generates synthetic speech in the *target language* from the translated text.

In this paper, we present the architecture of *Rhonda*[1], a scalable and multilingual capable S2S translation pipeline. Rhonda is agnostic in terms of the underlying HLT components and in fact can accommodate multiple implementations of the underlying technologies (e.g. multiple and differing TTS engines) simultaneously in order to:

- Improve multilingualism: there might be different implementations of the HLT components for different languages.
- Provide a unified interface: a unified interface is provided to calling applications through polymorphism of the underlying HLT components.

Although the various required and optional HLT components in an S2S translation system are discussed in the following sections, the focus of this paper is on the design, architecture and implementation of the S2S translation pipeline that controls the flow of information between the components and the calling applications.

The organisation of the paper is as follows: Section II details previous work on S2S translation systems and their HLT components, whilst Section III presents the design and architecture of the Rhonda S2S translation pipeline. Section IV describes the implementation details of the presented system. Lastly, a discussion and conclusion is given in Section V.

II. Background and Related Work

One of the first known speech-to-speech translation systems was developed by the NEC Corporation[2] and demonstrated as

[1]Rhonda is an anagram of a Hadron, a subatomic composite particle made of quarks (denoting the HLT components). The anagram is a play on the fact that the HLT components can be piped together in any desired fashion.

[2]https://www.nec.com/

a concept exhibit at the 1983 International Telecommunication Union (ITU) Telecom World event [5]. The *Advanced Telecommunications Research Institute International* (ATR) in Japan began research into automated S2S translation in 1986, with it's first phase focused on a feasibility study of a limited vocabulary and clear read speech system, the second phase on conversations within a limited domain and their current focus on the development of S2S translation systems in real world environments [6].

Various joint research projects and consortiums have been established in order to further the state-of-the-art of automated S2S translation systems, these include:

- *Consortium for Speech Translation Advanced Research* (C-STAR): An international consortium including ATR in Japan, *Carnegie Mellon University* (CMU) in the United States, *the Institute for Research in Science and Technology* (IRST) in Italy, the *Chinese Academy of Sciences* (CAS), and the *Electronics and Telecommunications Research Institute* (ETRI) in Korea [5].
- *Universal Speech Translation Advanced Research* (U-STAR): An international research consortium currently comprised of 33 institutes from 26 countries or regions [7].
- *Asian Speech Translation Advanced Research* (A-STAR): An international consortium, consisting mostly of countries in the Asian region, including Japan, China, Korea, Thailand, Indonesia, India, Singapore and Vietnam [8].
- *Technology and Corpora for Speech to Speech Translation* (TC-STAR): A joint research project of the European Union under the 6$^{\text{th}}$ Framework Programme for Research and Development [9].
- *Vermobil*: A research and development project funded by the German Ministry for Research and Technology (BMFT) and an industrial consortium with international partners [10].

C-STAR started the *International Workshop on Spoken Language Translation* (IWSLT) in 2004, where spoken language translation technologies from participating parties are evaluated on a multilingual speech corpus containing the tourism-related sentences that have been developed by ATR and C-STAR members. The workshop focused on the evaluation and technical papers on spoken language translation technologies [11]. The workshop has been held every year since 2004.

The research and development activities on S2S translation systems have been concentrated on the underlying HLT components in the context of an S2S translation system as well as the architecture and development of the S2S translation system.

A. HLT components of an S2S translation system

The speech-to-speech translation task, through a statistical approach, can formally be given as follows [6]:

$$S_T^* = \arg\max_{S_T} P(S_T|S_S) \tag{1}$$

where S_S and S_T are the respective source language and target language speech signals. The conditional probability $P(S_T|S_S)$ can be factorized as:

$$\begin{aligned} P(S_T|S_S) &= \sum_{T_T,T_s} P(S_T,T_T,T_S|S_S) \\ &= \sum_{T_T,T_s} P(S_T|T_T,T_S,S_S)P(T_T|T_S,S_S)P(T_S|S_S) \\ &\approx \sum_{T_T,T_s} P(S_T|T_T)P(T_T,T_S)P(T_S|S_S) \end{aligned} \tag{2}$$

where T_S and T_T are the transcriptions of the respective source language and target language speech signals. The maximisation of $P(S_T|S_S)$ in (1) can then be simplified to:

$$\max_{S_T} P(S_T|S_S) = \max_{S_T} P(S_T|T_T) \max_{T_T} P(T_T|T_S) \max_{T_S} P(T_S|S_S) \tag{3}$$

This result suggests that the automatic S2S translation task can be decomposed into three *independent* tasks: $P(T_S|S_S)$, denoting automatic speech recognition; $P(T_T|T_S)$, representing machine translation; and $P(S_T|T_T)$, which is the text-to-speech component.

A counter argument to the separation of the S2S components are that in practice the individual components do not produce perfect results and that in a serial combination of these, the error multiplies. In [12], a finite state translation model is tightly integrated into the acoustic-phonetic model of the automatic speech recognition system, resulting in a target language transcription for the source language speech signal.

One of the performance metrics of an ASR system is the *word error rate*[3] (WER). The results in [13] suggests that optimising over the combination of the ASR-MT component to a performance metric oriented to machine translation, such as the *bilingual evaluation understudy*[4] (BLEU), the translation quality of the ASR-MT combination is improved, even thought the WER on the ASR component is sacrificed.

The source language transcription produced by the ASR component and passed on to the MT component is usually the single-best recognition hypothesis. [14] argues that an MT system's translation quality can be improved if it can utilise supplementary information available from the ASR system, such as: N-best recognition hypotheses and likelihoods of acoustic and language models.

In a multilingual environment, where the ASR component has the capability of recognising more than one source language, there needs to be some or other method to inform the ASR component of which source language to expect. This can be achieved either through an extra input parameter or through a spoken language identification system. In [15] a joint language identification and speech recognition method

[3]https://en.wikipedia.org/wiki/Word_error_rate
[4]https://en.wikipedia.org/wiki/BLEU

was proposed, this method generates a clasification decision on a partial hypothesis.

Human speech is a complex signal conveying a rich spectrum of information such as identity, emotion and state of the speaker, style, and intent through the choice of words [16]. Recent research on improving the naturalness of S2S translation systems have focused on emphasis [17], emotion [18] and punctuation and segmentation [19].

B. Architecture of an S2S translation system

In 2010 U-STAR standardized two international communication protocols, approved by the *Telecommunication Standardization Sector of ITU* (ITU-T), that define the *functional requirements for network-based speech-to-speech translation services* (ITU-T F.745) [20] and the *architecture for network-based speech-to-speech translation services* (ITU-T H.625) [21]. These two recommendations have since been updated in 2016 and 2017 respectively.

Many published works [6], [12], [22], [23] on S2S translation systems are closely integrated and do not go into the detail of the design of a network-based S2S translation system as defined in [20] and [21].

The Vermobil project [10] has a "modularized functional architecture" [24], with interactively communicating modules, where the functionality of the S2S translation system is broken down into separate acoustic and linguistic modules. The acoustic modules also include a prosody module, which annotates the word-lattices as recognized by the automatic speech recognition modules with prosodic information. The basic software architecture is described, but not in detail concerning the interface protocols of the modules.

The ITU-T F.745 recommendation specifies the "service description and the requirements for speech-to-speech translation" through a network as well as the components of a S2S translation system as follows [20]:

- *S2S client*:
  - user client for speech/text input and output.
- *S2S servers*:
  - speech recognition: speech is recognized and transcribed;
  - machine translation: text in source language is translated into text in target language;
  - speech synthesis: speech signal is created from text.
- *Communication protocol*:
  - communication protocol to connect user clients and the above S2S servers.

whilst the ITU-T H.625 recommendation defines the following items for the design of a network-based S2S translation system [21]:

- *a functional architecture and mechanisms of network-based S2S translation system;*
- *interface protocols between S2S translation modules; and*
- *a workflow of the network-based S2S translation system.*

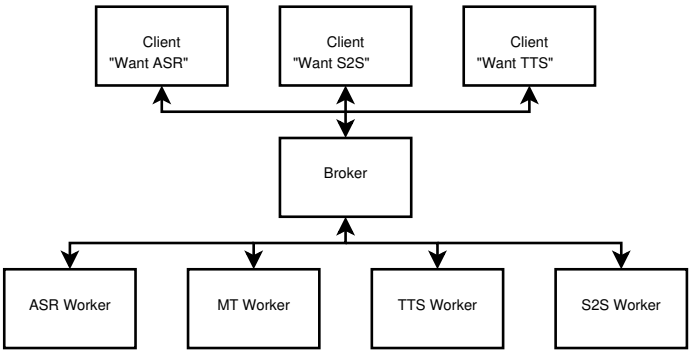


Fig. 1. The functional architecture of the Rhonda S2S translation pipeline.

In [25] an implementation of the U-STAR specification is described in detail. The application is a multi-device application, named "VoiceTra4U-M", available on the iPhone and Android platforms. It allows up to 5 users to communicate in different languages, either face-to-face or remotely.

III. Design and Architecture

The design requirements of the Rhonda speech-to-speech translation pipeline can be summarised as follows:

- *Reliability*: the requirement that a S2S request is always serviced, irrespective of individual HLT components failing.
- *Scalability*: the requirement that the S2S pipeline service can scale on demand and as required (dynamically adding more HLT components as required).
- *Distributed*: the requirement that the HLT components can be deployed on distributed hardware.

Although all these requirements talk to a network-based S2S translation system as defined in [21], the choice of the interface protocols between the S2S translation modules dictates whether the system as a whole will be able to function on-device and off-line, as the network may be implemented with an *inter-process communication*[5] (IPC) protocol (the device hardware restrictions of course also play a role). The scalability and distributed design requirements would also fall away for an on-device and off-line implementation although the aim is to create a design and architecture that will be portable and have on-device and off-line capability.

The following subsections address the design of the Rhonda S2S translation pipeline in terms of the ITU-T H.625 recommendations [21].

A. Functional Architecture

The functional architecture of the Rhonda S2S translation pipeline is shown in Figure 1. The system is a service-oriented request-reply dialog between a set of client applications, a broker, and a set of worker applications. The architecture is based on the *Majordomo Protocol* [26] (MDP), with added elements of the *Majordomo Management Interface* [27] (MMI). There

[5]https://en.wikipedia.org/wiki/Inter-process_communication

are three actors in the Rhonda S2S translation system, which at the most basic level can be described as:

- *Clients*: request HLT services;
- *Workers*: supply HLT services; and
- *Broker*: connects clients which request a specific HLT service to workers which supply those HLT services.

The Rhonda functional mechanisms define the interaction between these three actors. In Figure 1 there are three clients, each requesting a different service, "ASR", "S2S" and "TTS". There are also three workers implementing those services, as well as a worker implementing an "MT" service. The *clients* in Rhonda correspond to the S2S translation clients as defined in [20], whilst the *workers* correspond to the S2S translation servers in [20]. The Rhonda *broker* handles the following:

- Communication between the clients and workers;
- The status of workers; and
- New workers coming online.

Neither the number of clients or workers are limited. The workers may implement any HLT component as a service and the clients may request any HLT service as implemented by a worker, or even a list of HLT services. Multiple workers may implement the same HLT component as a service (i.e. automatic speech recognition), where the component may be the same underlying technology (the same ASR engine) or a differing instance of the technology (i.e. a different type of ASR engine). A basic assumption about the workers are that they are *idempotent*, in other words they do not store any state and the same HLT service request will result in the same response, irrespective of the client or any other external factors. Another assumption about the workers is that they can handle at most one request a time, and will issue exactly one reply for each successful request.

B. Interface Protocols

The Rhonda broker handles the interface protocols between the clients and workers. The protocol which is based on [26] and [27] can decomposed into two sub-protocols, the client and worker protocols.

1) Client Protocol: The client protocol is a synchronous request-reply dialog between the client and the broker, initiated by the client, and can be represented as follows:

```
Repeat:
    Client: REQUEST
    Broker: REPLY
    ...
```

A `REQUEST` command consists of a multipart message, containing the following:

- The protocol version;
- The requested service; and
- The request body.

The `REPLY` command is also a multipart message, containing the following:

- The protocol version;
- The requested service; and
- The reply body.

Clients use a polling strategy in order to recover from a non-responsive broker. If there is no reply within some timeout, then the connection between the client and broker is closed and a new connection is opened where the `REQUEST` command is resent. If there is no reply after several retries, then the client is to signal that transaction as failed to the calling application.

2) Worker Protocol: The worker protocol is a mix of a synchronous request-reply dialog between the worker and the broker, initiated by the worker, and an asynchronous "heartbeat"[6] dialog that operates independently in both directions. The synchronous request-reply dialog can be represented as follows:

```
Worker: READY
Repeat:
    Broker: REQUEST
    Worker: REPLY
    ...
```

A `READY` command by the *worker* consists of a multipart message, containing the following:

- The protocol version; and
- The HLT service implemented by the worker.

A `REQUEST` command by the *broker* also consists of a multipart message, containing the following:

- The protocol version;
- The client address (of the client requesting the service); and
- The request body.

The `REPLY` command by the *worker* is a multipart message, containing the following:

- The protocol version;
- The client address (of the client which requested the service); and
- The reply body.

The asynchronous "heartbeat" dialog can be represented as follows:

```
Repeat:                      Repeat:
    Worker: HEARTBEAT            Broker: HEARTBEAT
    ...                          ...
Worker: DISCONNECT           Broker: DISCONNECT
```

A `HEARTBEAT` command by the *broker and worker* consists of a multipart message, containing the following:

- The protocol version; and
- A heartbeat message.

A `DISCONNECT` command by the *broker and worker* also consists of a multipart message, containing the following:

- The protocol version;
- A disconnect message.

The interaction between the broker and the workers is as follows:

[6] https://en.wikipedia.org/wiki/Heartbeat_(computing)

- The worker initiates the interaction between the worker and the broker and registers itself and it's HLT service with the broker with the `READY` command.
- The worker will send a periodic `HEARTBEAT` command to the broker in order to determine if the broker is still responsive and the worker *must* also respond to periodic `HEARTBEAT` commands received from the broker.
- If either the broker or the worker does not receive a response from a `HEARTBEAT` command, then they send `DISCONNECT` commands and try to reconnect.
- When a worker receives a `REQUEST` from the broker it acts on the HLT service request and forms a `REPLY` and sends it to the broker, which then sends it along to the client which requested the service.

3) Reply/Request Data Structure: In order to allow future expansion to include any type of HLT component, as either a service or a component used to improve automated S2S translations, the data structure of the `REQUEST` and `REPLY` body of Sections III-B1 and III-B2 need to be defined with the heterogeneous outputs of HLT components in mind. Previous research in TTS systems have resulted in two widely used representations [28], namely the *delta system* [29] and *Heterogeneous relation graphs* [30] (HRG).

Both the delta system and HRGs are top level "containers", which allow building data structures based on linguistic items within these containers. Items can be any linguistic unit, from a word, phone, syllable or prosodic description to a sentence or a list of sentences. Each item may also have attributes or features that are associated with it. These items can then be combined in arbitrary fashions in order to add structure to a global linguistic item in terms of *relations*.

Figure 2 shows an example representation of an utterance structure, for the text input "twenty fifth", using a HRG with four relations and their items. The relations are:

- "Word": where the items are the words in the utterance.
- "Syllable": the items are the syllables in the words.
- "Segment": the items represent phones in the words.
- "SylStruct": a relation that connects the "Word", "Syllable" and "Segment" relations.

Note that for brevity the phones of the word "fifth" have been omitted. The figure also shows the item's features, for example, the "Word" relation's items have as a "name" feature the word it was derived from, whereas the "Syllable" relation's items have a "stress" feature (0 or 1). The number of features are conceptually unlimited, the only restriction being that the feature names are unique in a specific item.

The delta system is not as powerful in it's representations as the HRG as it cannot represent general tree information [30], and therefore the HRG data structure has been chosen to represent the `REQUEST` and `REPLY` bodies of the Rhonda S2S pipeline interface protocols. The HRG representation is also widely used in TTS systems [31]–[33] for internal data representation. It must be noted that HRGs are not limited to linguistic information, and can also contain any signals associated with human speech.

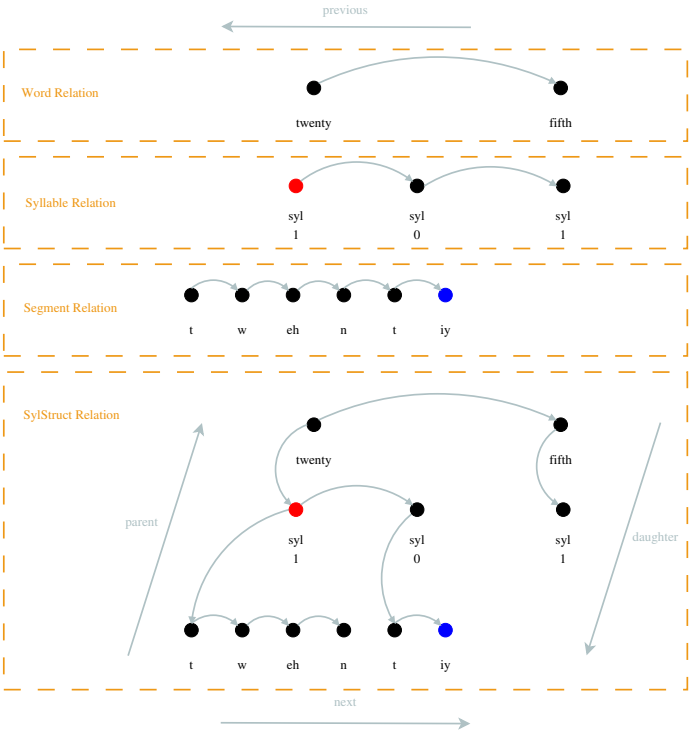


Fig. 2. An example representation of an utterance structure using a heterogeneous relation graph (HRG).

4) API Protocol: One of the goals of the Rhonda S2S translation pipeline is to support any HLT component and technology. Each HLT (i.e. machine translation or text-to-speech) or even instance of HLT (i.e. two different automatic speech recognition engines) may have a different *application programming interface*[7] (API). This places the responsibility of publishing the API on the worker which implements the HLT component, for consumption by the client requesting the HLT service. This is accomplished by using the *OpenAPI Specification* [34]. The OpenAPI specification was originally developed as an API specification for machine-readable interface files for describing, producing, consuming, and visualizing RESTful [35] web services. In this work it is used for describing the HLT component implementation API by the Rhonda worker and consumption of the API by the Rhonda client. This allows a client to query the API in order to determine if the required service is implemented by a specific worker.

C. Workflow

The workers implement an HLT component as a service in the Rhonda S2S translation pipeline, bridging the Rhonda protocols and API and the API of the specific HLT component. The clients bridge the API of the calling application (which may be a RESTful web service or any other application) and the Rhonda protocols and API. Only the broker needs to be known between the clients and the workers.

[7]https://en.wikipedia.org/wiki/Application_programming_interface

Workers connect to the broker and publish their API through the OpenAPI specification with the protocol as mentioned in Section III-B2. Clients connect to the broker and request an HLT service with the protocol as mentioned in Section III-B1. Clients can also request a list of services that are provided by Rhonda and query the API of the services.

When the broker receives a request from the client it constructs the request into an HRG data structure and passes it to the first available worker implementing the requested service (if there is more than one worker implementing the service then the broker will use the last used worker).

The worker receives the request from the broker in an HRG data structure and replies with an HRG data structure containing the fulfilled request. The broker then sends this reply back to the requesting client.

IV. Implementation

The current implementation of the clients, workers and broker of the Rhonda S2S translation pipeline has been implemented in the Python[8] programming language. The communication protocols as mentioned in Section III-B have been implemented using the ZeroMQ[9] libraries.

ZeroMQ is an embeddable networking library that provides sockets[10] across various transports, which include:

- in-process;
- inter-process;
- Transmission control protocol[11] (TCP); and
- multicast[12].

ZeroMQ allows for various N-to-N connection patterns and it's API is accessible in a multitude of programming languages (C, PHP, Java, Python, Lua, Haxe, C++, C#, CL, Delphi, Erlang, F#, Felix, Haskell, Objective-C, Ruby, Ada, Basic, Clojure, Go, Haxe, Node.js, ooc, Perl, and Scala). This provides many future options for porting of the Rhonda clients, workers and broker to other programming languages.

Worker implementations have been interfaced with the *PocketSphinx* [36] ASR engine, a *Grammatical framework* based MT system and the *Speect* [33] TTS engine. This string of HLT components allows for a initial baseline of the full functionality Rhonda S2S translation pipeline.

A client implementation has been written, providing a RESTful API to the above mentioned worker implementations through the *Tornado*[13] web framework and accessible to any application that can do a RESTful API request.

V. Discussion and Conclusion

In this paper we have presented a new architecture and implementation of a speech-to-speech translation pipeline named Rhonda. Our main design goals were given in Section III, and were focused around reliability, scalability and distributed.

[8]https://www.python.org/

[9]http://zeromq.org/

[10]https://en.wikipedia.org/wiki/Network_socket

[11]https://en.wikipedia.org/wiki/Transmission_Control_Protocol

[12]https://en.wikipedia.org/wiki/Multicast

[13]https://www.tornadoweb.org

The reliability aspect of the system has been achieved by the Rhonda constantly monitoring the status of the HLT component implementations. This allows for the option of additional service monitoring systems to be used to shutdown and startup worker implementations as is needed.

The scalability requirement has been achieved by the design and architecture of Rhonda, where multiple instances of an HLT component can be implemented by a Rhonda worker. Future work will focus on adding load monitoring systems, which can dynamically startup and shutdown HLT worker implementations as is required.

Finally, the distributed requirement has been met by the choice of the interface protocols and the choice of the implementation libraries. ZeroMQ allows for the Rhonda clients, workers and brokers to be distributed on different hardware, with only the address of the broker to be known to the clients and the workers. Therefore, one may make use of additional hardware in order to scale the pipeline as is required by the load.

We have also integrated text-based language identification (LID) and named entity recognition (NER) worker implementations into Rhonda, proving that the pipeline can be utilised as a general HLT pipeline as well.

In initial tests, with dummy client and worker implementations, the pipeline was able to handle around 5000-6000 client requests per second on a single machine. Our future work will include stress testing the pipeline with multiple external clients and varying the load over time in order to determine the hardware requirements for scaling versus the number of simultaneous connections.

The next steps in our development include interfacing Rhonda into a real-world application, being developed on a mobile platform. The piloting of this application will guide further development and enhancements to the Rhonda pipeline.

References

[1] E. A. Jacobs, D. S. Shepard, J. A. Suaya, and E.-L. Stone, "Overcoming language barriers in health care: costs and benefits of interpreter services," *American journal of public health*, vol. 94, no. 5, pp. 866–869, 2004.

[2] L. Herbst and S. Walker, "Language barriers in the delivery of police services: A study of police and hispanic interactions in a midwestern city," *Journal of Criminal Justice*, vol. 29, no. 4, pp. 329–340, 2001.

[3] R. C. Davis, E. Erez, and N. Avitabile, "Access to justice for immigrants who are victimized: The perspectives of police and prosecutors," *Criminal Justice Policy Review*, vol. 12, no. 3, pp. 183–196, 2001.

[4] H. Tenzer, M. Pudelko, and A.-W. Harzing, "The impact of language barriers on trust formation in multinational teams," *Journal of International Business Studies*, vol. 45, no. 5, pp. 508–535, 2014.

[5] S. Nakamura, "Overcoming the Language Barrier with Speech Translation Technology," *NISTEP Quarterly Review*, no. 31, pp. 35–48, 2009.

[6] S. Nakamura, K. Markov, H. Nakaiwa, G. ichiro Kikui, H. Kawai, T. Jitsuhiro, J.-S. Zhang, H. Yamamoto, E. Sumita, and S. Yamamoto, "The ATR Multilingual Speech-to-Speech Translation System," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 365–376, 2006.

[7] (2018, jun) U-STAR - The Universal Speech Translation Advanced Research Consortium. [Online]. Available: http://www.ustar-consortium.com/qws/slot/u50227/index.html

[8] S. S. Watiasri, M. Paul, A. Finch, S. Sakai, T. T. Vu, N. Kimura, C. Hori, E. Sumita, S. Nakamura, J. Park, C. Wutiwiwatchai, B. Xu, H. Riza, K. Arora, C. M. Luong, and H. Li, "A-STAR: Toward Translating Asian Spoken Languages," *Computer Speech & Language*, vol. 27, no. 2, pp. 509–527, 2011.

[9] G. Lazzari and V. Steinbiss, "Human language technologies for europe," *ITC IRST/TC-Star project report*, 2006.

[10] W. Wahlster, "Verbmobil," in *Grundlagen und anwendungen der künstlichen intelligenz*. Springer, 1993, pp. 393–402.

[11] (2018, jun) International Workshop on Spoken Language Translation (IWSLT) 2004. [Online]. Available: https://www.isca-speech.org/archive/iswlt_04/

[12] E. Vidal, "Finite-state speech-to-speech translation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Munich, Germany, April 1997, pp. 111–114.

[13] X. He, L. Deng, and A. Acero, "Why word error rate is not a good metric for speech recognizer training for the speech translation task?" in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011, pp. 5632–5635.

[14] R. Zhang, G. Kikui, H. Yamamoto, F. Soong, T. Watnabe, and W. K. Lo, "A Unified Approach in Speech-to-Speech Translation: Integrating Features of Speech Recognition and Machine Translation," in *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, 2004.

[15] D. C. Y. Lim, I. Lane, and A. Waibel, "Real-Time Spoken Language Identification and Recognition for Speech-to-Speech Translation," in *International Workshop on Spoken Language Translation (IWSLT)*, Paris, France, 2010, pp. 307–312.

[16] S. Johar, *Emotion, affect and personality in speech: The Bias of language and paralanguage*. Springer, 2015.

[17] Q. T. Do, S. Takamichi, S. Sakti, G. Neubig, T. Toda, and S. Nakamura, "Preserving word-level emphasis in speech-to-speech translation using linear regression HSMMs," in *Sixteenth Annual Conference of the International Speech Communication Association*, Dresden, Germany, 2015.

[18] M. Akagi, X. Han, R. Elbarougy, Y. Hamada, and J. Li, "Toward affective speech-to-speech translation: Strategy for emotional speech recognition and synthesis in multiple languages," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), Asia-Pacific*, Chiang Mai, Thailand, 2014, pp. 1–10.

[19] E. Cho, J. Niehues, and A. Waibel, "Domain-independent Punctuation and Segmentation Insertion," in *International Workshop on Spoken Language Translation (IWSLT)*, Tokyo, Japan, 2017, pp. 90–97.

[20] T. S. S. of ITU, "Functional requirements for network-based speech-to-speech translation services," International Telecommunication Union (ITU), Recommendation F.745, July 2016. [Online]. Available: https://www.itu.int/rec/T-REC-F.745/en

[21] ——, "Architecture for network-based speech-to-speech translation services," International Telecommunication Union (ITU), Recommendation H.625, March 2017. [Online]. Available: https://www.itu.int/rec/T-REC-H.625/en

[22] A. Lavie, A. Waibel, L. Levin, M. Finke, D. Gates, M. Gavalda, T. Zeppenfeld, and P. Zhan, "JANUS-III: Speech-to-speech translation in multiple languages," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Munich, Germany, April 1997, pp. 99–102.

[23] A. Waibel, A. N. Jain, A. E. McNair, H. Saito, A. G. Hauptmann, and J. Tebelskis, "JANUS: a speech-to-speech translation system using connectionist and symbolic processing strategies," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Ontario, Canada, May 1991, pp. 793–796.

[24] T. Bub and J. Schwinn, "Verbmobil: The evolution of a complex large speech-to-speech translation system," in *The 4th International Conference on Spoken Language Processing (ICSLP)*, Philadelphia, USA, October 1996, pp. 2371–2374.

[25] K. Arora, S. Arora, and M. K. Roy, "Speech to speech translation: a communication boon," *CSI Transactions on ICT*, vol. 1, no. 3, pp. 207–213, Sep 2013. [Online]. Available: https://doi.org/10.1007/s40012-013-0014-4

[26] P. Hintjens, C. Remes, and G. Goldstein, "Majordomo protocol," iMatix Corporation, Tech. Rep. 7/MDP, 2011. [Online]. Available: https://rfc.zeromq.org/spec:7/MDP

[27] P. Hintjens, "Majordomo management interface," iMatix Corporation, Tech. Rep. 8/MMI, 2011. [Online]. Available: https://rfc.zeromq.org/spec:8/MMI

[28] P. Taylor, *Text-to-Speech Synthesis*. Cambridge University Press, 2009.

[29] S. R. Hertz, "The delta programming language: an integrated approach to non-linear phonology, phonetics and speech synthesis," in *Papers in Laboratory Phonology: Volume 1, Between the Grammar and Physics of Speech*, J. Kingston and M. E. Beckman, Eds. Cambridge University Press, 1990, vol. 1, p. 215.

[30] P. Taylor, A. W. Black, and R. Caley, "Heterogeneous relation graphs as a formalism for representing linguistic information," *Speech Communication*, vol. 33, no. 1-2, pp. 153–174, 2001.

[31] ——, "The architecture of the Festival speech synthesis system," in *3rd ESCA Workshop on Speech Synthesis*. Jenolan Caves, Australia: International Speech Communication Association, 1998, pp. 147–151.

[32] A. W. Black and K. A. Lenzo, "Flite: a small fast run-time synthesis engine," in *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*, Perthshire, Scotland, 2001, pp. 157–162.

[33] J. A. Louw, "Speect: a multilingual text-to-speech system," in *Proceedings of the Nineteenth Annual Symposium of the Pattern Recognition Association of South Africa*, Cape Town, South Africa, November 2008, pp. 165–168.

[34] T. O. Initiative, "Openapi specification," The OpenAPI Initiative, Tech. Rep. Version 2.0, 2014. [Online]. Available: https://swagger.io/specification/v2/

[35] R. T. Fielding, "Rest: architectural styles and the design of network-based software architectures," Ph.D. dissertation, University of California, 2000.

[36] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. I. Rudnicky, "Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, May 2006, pp. I–I.

Design of an Autonomous Management and Orchestration for Fog Computing

Sabelo Dlamini
CSIR Meraka Institute
Pretoria, South Africa
sabelo.dlamini@csir.co.za

Joyce Mwangama, Neco Ventura
University of Cape Town
Cape Town, South Africa
{joycebm, neco}@crg.ee.uct.ac.za

Thomas Magedanz
Fraunhofer FOKUS Institute
Berlin, Germany
thomas.magedanz@fokus.fraunhofer.de

***Abstract*—Fog computing aims to bring cloud computing capabilities to the edge of the network, closer to the end user, enabling lower latencies, location awareness, and mobility support among other advantages. The combination of IoT and Fog encompasses a highly complex scenario with a huge amount of data and varying number of devices that must cooperate with each other. Fog computing networks may be designed as autonomous networks. In this case, there is a requirement for effective management and orchestration mechanisms to guarantee acceptable performance of applications and services, while still leveraging of cloud capabilities. Mechanisms typically applied to "cloud-only" implementations cannot naturally be migrated to the Fog given its particular characteristics. This calls for the design and development of new management and orchestration mechanisms for the Fog. In this paper we propose a design the use the finite state machine to enhance decision making in an autonomous Fog computing network. The proposed scheme is expected to optimise Fog computing networks autonomy and improve performance and cost.**

Keywords—Fog Computing, IoT, ETSI NFV MANO, OpenFog, 5G

I. INTRODUCTION

The growth of the Internet of Things (IoT) is driving business transformation by connecting everyday objects and devices together and to cloud-hosted services. It is estimated that about 28 billion smart devices will be connected across the world by 2021, with more than 15 billion of these devices to be connected through M2M and consumer-electronics devices [1]. This has prompted the need to provide networking, computing and storage capabilities closer to the end user in order to meet the stringent requirements of low latency and high bandwidth for IoT applications.

Several technologies have been designed in response to this expansion of IoT devices, such as network function virtualization, fifth generation networks and fog computing. In particular, the European Telecommunications Standards Institute (ETSI) has standardized the reference architecture for NFV management and orchestration (MANO) [2], a cornerstone for building, deploying, and managing services in NFV environments. Advances in the 5G radio access network (RAN) and Multi-access Edge Computing (MEC) group at ETSI [3]. The adoption of the OpenFog reference architecture by IEEE are also key for the IoT evolution [4].

This increase in IoT devices means the current "cloud-only" architectures cannot keep up with the volume and velocity of this data across the network, thereby reducing the value that can be created and captured from these investments [4]. In order to address the challenges and limitations of the "cloud-only" architecture, Fog computing provides the missing link in the cloud-to-thing continuum. Fog computing architectures selectively move compute, storage, communication, control, and decision making closer to the network edge where data is being generated in an effort to address the limitations of the current infrastructure to enable mission-critical and data-dense use cases [4].

Fog computing is an extension of the traditional cloud-based computing model where implementations of the architecture can reside in multiple layers of a network's topology. However, all the benefits of the cloud should be preserved with the extension to Fog, including containerisation, virtualisation, orchestration, manageability, and efficiency. In many cases, Fog computing works with the cloud. In this work we define Fog computing as an independent extension of the cloud.

Fog computing also enhances network performance in case of latency sensitive IoT applications which are difficult to provide over cloud computing, because of limitations such as latency, bandwidth and cost of the backhaul networks. Fog computing provides the channel which is proxy-like to bring services closer to the IoT devices thereby reducing traffic overhead on the backhaul and decreasing propagation delays.

In addition to network performance, as a technological paradigm, fog computing networks may be architecturally organized as an autonomic (self-healing) computing network. This has been highlighted as a key requirement in the OpenFog reference architecture [4]. This autonomy helps in cases where data generated by IoT devices cannot be cost-effectively sent to the cloud, or there may be legal requirements or security considerations not to send it to the cloud. In these cases, Fog computing provides a way of alternative local capabilities of network, compute and storage.

In this paper we propose an enhancement to autonomous management and orchestration capabilities of Fog computing networks by adopting the ETSI NFV MANO framework with finite state machine to enhance autonomous decision making. Finite state machine is a mathematical model of computation. It

is an abstract machine that can be in one state at any given moment. The finite state machine can change from one state to another in response to external inputs. We believe that finite state machines are efficient and easier to implement in a design compared to other artificial intelligence implementations such as machine learning, which require data collection and data analysis for the performance of the network [5].

The reminder of the paper is organized as follows: Section II provides an overview of related work in Fog computing and management and orchestration, while Section III describes the proposed scheme architecture. Section IV describes the design and validation of the proposed scheme, and Section V describes the initial evaluations, while in Section VI the conclusion and future work discussion are provided.

II. Related Work

Autonomous management and orchestration have been a key research challenge in Fog computing. Several research projects have proposed different autonomous management and orchestration schemes. The most common trend is the integration of different architectures in addressing Fog computing challenges. van Lingen in [6] highlighted the need for convergence of NFV, 5G and Fog to address these challenges. They highlight the importance of interplay between cloud and fog to cater for the evolution and demand of IoT. They identified ETSI NFV MANO as the key architecture to enable this, they further propose a scheme that uses it to provide uniform management of IoT services.

Reference [7] proposed a converged architecture that brings together the strengths of OpenFog and ETSI NFV MANO architectures and applies them in an IoT space. While Santos et al. in [8] extended the ETSI NFV MANO framework for Fog computing paradigms by introducing a fully-integrated and autonomous Fog node management system function, which combines monitoring and data analysis operations alongside management and orchestration decisions.

The research work by José Santos et al. in [8] proposed a framework for enabling autonomous management and orchestration functionalities in smart city edge networks. José Santos et al. uses different sets of machine learning algorithms to provide self-management responses to unusual events or malfunctions that can be detected in the network. Their proposed framework then applies specific network strategies corresponding to the detected malfunctions from the policy catalog.

Also, de Brito et al. in [9] highlights the challenge of orchestration in the edge computing environment and proposes an infrastructure management and orchestration architecture for an edge computing environment. The proposed architecture switches between the centralized and distributed model. The switching between models is both manually and automatically started through event triggers in the framework. The information on how the automatic switching was not provided.

The SELFNET project in [10] proposes a novel approach to autonomously deal with network failures, this architecture uses intelligent algorithms such as feature selection and classification to develop network intelligence which enables an autonomic management for 5G networks. However, the proposed scheme/solution was for 5G mobile networks, as such the solution is more resource intensive which makes it difficult to be implemented in edge networks.

The CogNet project in [11] applies autonomic network management based on Machine Learning as a key technology for 5G networks to reach the vision of automated management of telecoms network infrastructures. Cognitive Smart Engine is used in the proposed scheme to provide near real-time learning, in order to dynamically adapt resources to the immediate requirements of the virtual network functions, while minimizing performance degradations to fulfill service level agreement (SLA) requirements.

In this paper we propose the use of finite state machine for modelling of an autonomous edge network which can be used to enhance decision making in management and orchestration. To the best of our knowledge, there has been not related work that has explored the use of finite state machine for management and orchestration. As outlined above in the related work, most of the work has explored different implementation of artificial intelligence such machine learning to enhance decision making to management and orchestration. Most machine learning implementations require training process which is often time consuming and involves high cost [5], therefore we will believe the use of finite state machine will offer a more efficient method for management and orchestration.

The finite state machine has been successfully implemented by the most widely used communication protocols. The known protocols that are modelled using finite state machines are Border Gateway Protocol (BGP) and Controller Area Network (CAN) [12]. They have also been adopted in even more complex scenarios such in mobile robot's navigation to provide autonomous intelligence capability [13].

III. Proposed Architecture

In the design of the proposed scheme we are going to adopt OpenFog and ETSI NFV MANO architectures, in addition to this we will add the finite state machine component to enhance decision making in the edge node of the network.

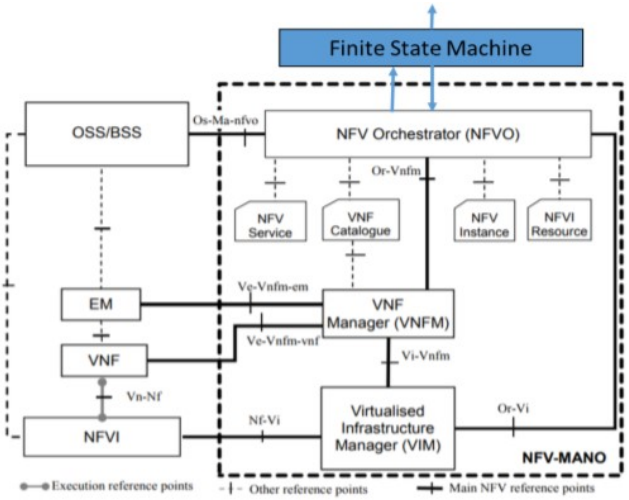


Figure 1: ETSI NFV MANO with Finite State Machine Module

The finite machine module will receive the status of the Fog network and execute event triggers based on the state of the network. The proposed scheme will be sitting at the edge of the autonomous network as shown in Figure 2. It will monitor the state of the autonomous network. If the network performance indicators are not satisfying the network user requirements, the proposed scheme will automatically change the state of the network. Performance metrics such as latency, bandwidth and cost will serve as inputs to the decision making.

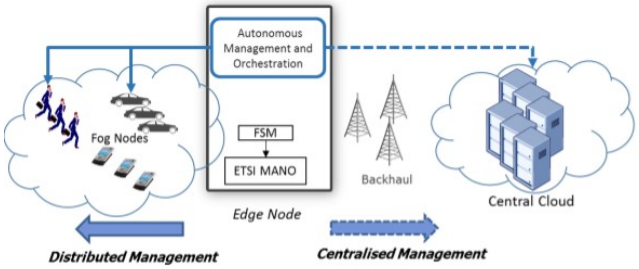


Figure 2: Autonomous Management and Orchestration Architecture

The proposed scheme will use latency, bandwidth, energy efficiency and cost as inputs. Cost of the network will be defined as the use of backhaul network. If these fall below defined thresholds for services' requirements the autonomous network will change its state as discussed in the scheme design in the next section.

The network above is modelled into 3 states, independent autonomous network, integrated edge with unreliable backhaul and end-to-end integrate edge. These are discussed in detail below:

1) Independent Autonomous Network: This is the default state where the network is fully autonomous and is able to independently make decisions and function on its own. In this state the storage, computation and networking are locally implemented, the goal is to have the edge network spending most of its time in this state. In this state the network is able to handle routine tasks without integration with the cloud. Specific event triggers such as where the resources are insufficient will prompt the move to the next states discussed below.

2) Integrated edge with Unreliable backhaul: This is the state whereby the network has intermittent or sporadic connectivity to the cloud at best effort. In the integrated edge with unreliable backhaul the networks needs to switch between centralized and distributed mode. In this state the storage and computation are locally implemented while networking is remote. The backhaul is best effort, the security low and there is higher latency, but the cost low. The network transition to this state if there is a requirement for best effort data backup, or data to be sent to the cloud for advanced computing.

3) End-to-end Integrated Edge: This is the state where the edge is fully integrated with the cloud. In this state the computation, storage and networking are remotely implemented. The computation is more powerful and the storage is sufficient. The latency is higher, because of the distance between end user and the cloud. The main disadvantage is that the cost of being in this state is high. In this state the edge network has full access to cloud advanced capabilities such as storage and compute with reliable networking capability to the cloud.

The end-to-end integrated edge state is the best state in terms of computing and storage metrics, but because of the location of the cloud there may be higher latency and cost of using backhaul network in addition to the cost of cloud services. The design goal is to have the network spend more time at the independent autonomous state and transition to other state when required. In the next section we will provide more details on how the finite state machine model is designed that will integrate with the ETSI NFV MANO to enable the network to transition between these states.

IV. Design and Implementation Method

There are different implementations of finite state machines based on different design requirements. In this paper will use the traditional finite state machine, with minimal past data and computation power requirements. This section will discuss the finite state modeling based on the proposed architecture and the details on the execution of event triggers for state transitions.

A. Finite State Machine Modelling

A finite-state automaton M = (S, E, e_0, s_0) consists of a finite set of states denoted by S, a finite set of events denoted by E, a transition event function e that assigns a next state to every pair of state and input (so that e_0: $S \times E \rightarrow S$), an initial or start state s_0.

The finite state machine will enable the fog network to switch between the different states in relations to cloud connectivity as shown in Table 1.

Table 1: State Definition for an Autonomous Edge Network

State	Definition
S_0	*independent autonomous network*
S_1	*integrated edge with unreliable backhaul*
S_2	*end-to-end integrated edge*

As shown in Table 1, the different states are defined as discussed in the previous section. The even triggers that will prompt the transition between the states will be discussed in the next section.

B. State Event Triggers

The state of the network is triggered by the change in the input metrics which are defined by x, y, z and t are input metrics latency, bandwidth and cost respectively, where t stands for time of the day and behaves as an overwrite condition for connecting to the cloud. The conditions are grouped into events which are defined in Table 2. The table outlines the event triggers associated with the edge network that enables transition between different states. These events are defined in high level, they are a combination of lack of demand for low latency, demand for higher bandwidth and reduction of backhaul network usage which is translated as cost in this study.

TABLE 2: EVENTS FOR THE AUTONOMOUS EDGE NETWORK

Event	Description
e_0	*Data Backup*
e_1	*Advanced Computing*
e_2	*Strenuous Computing*
e_3	*Daily Tasks*
e_4	*Results of computing obtained*
e_5	*Back up Complete*
e_6	*Storage Requirements Exceeded*

The event, e_0 is data backup this event is mostly informed by the time of the day, while e_6 event is triggered when the local storage is reaching full capacity. Event e_1 is advance computing this event is triggered when there is data to be sent to the cloud for computing. Strenuous computing e_2 event is the same as e_1. Daily tasks e_3 event returns the network to the independent state for daily network operations. Events e_4 and e_5 also prompt the network to revert the independent state.

C. State Transition Conditions

The state of the network is triggered by the change in the input metrics which are defined by *x, y, z* and *t* are input metrics latency, bandwidth and cost respectively, where *t* stands for time of the day and behaves as an overwrite condition for connecting to the cloud.

TABLE 3: STATE TRANSITION CONDITIONS FOR THE AUTONOMOUS EDGE NETWORK

	Inputs						Outputs	
Current State	**Event Triggers**						**Next State**	
	b_1	b_0	x	y	z	t	n_1	n_0
S_0	0	0	0	0	0	0	0	0
	0	0	0	0	0	1	1	0
	0	0	0	0	1	0	0	1
	0	0	0	0	1	1	1	0
	0	0	0	1	0	X	1	0
	0	0	0	1	1	X	0	0
	0	0	1	1	0	X	0	1
	0	0	1	1	1	X	0	0
S_1	0	1	1	0	0	X	0	0
	0	1	1	1	0	X	1	0
S_2	1	0	1	0	0	X	0	0
	1	0	1	0	1	X	0	0

The events, metrics and transition conditions are outlined in detail in Table 3. All combinations of *x, y, z* not listed for each state will by default remain in that current state. Table 3 outlines the state transitions where *current state* determines the current state the edge network is in defined by b_1 and b_0. The state then goes through specific network triggers and is prompted to move to the *next state* defined by n_1 and n_0 which both symbols correspond to the states defined in Table 1.

D. State Transition Model

Figure 2 summarizes the designed finite state machine model showing the transition between the states. The designed finite state machine model will use latency, bandwidth, and cost as inputs, combination of these metrics will define events that will necessitate state transitions as shown. If defined metrics fall below defined thresholds for services' requirements of the autonomous system an event trigger will be generated to change network state.

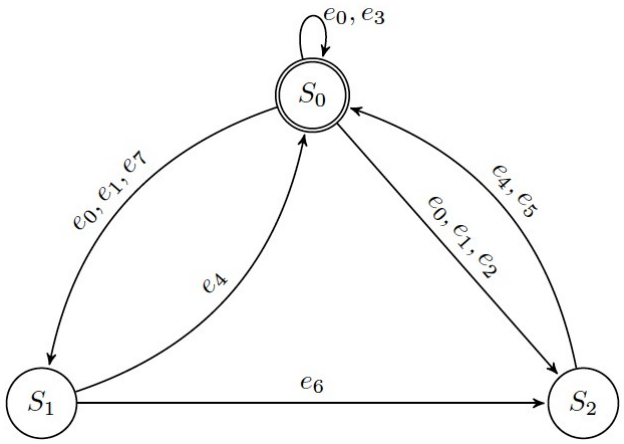


Figure 3: Proposed Schemes' State Transition Model

The state machine was designed, implemented and tested using MATLAB Simulink Stateflow. Stateflow is a graphical modeling and simulation environment, where each basic element of finite state machine theory can be displayed to build logic transformation model and event-driven system model. Stateflow, which is embedded into Simulink can dynamically control the state changes of target in real systems. The designed finite state machine was exclusive for all input combinations in all states.

Figure 3 shows the finite state machine from the Stateflow highlighting different combinations of *x, y, z* and *t* that initiates the events for the network to transition from one state to another. As shown having low latency and high bandwidth means *x=1, y=1, z=0* no matter what time of the day it is (*t=0*) creates and event trigger for a state to transition from S_0 to S_1. All other state transitions follow similar approach shown in the figure, in line with Table 3.

In this section we have outlined the proposed finite state machine was designed and validated on MATLAB Simulink Stateflow, and how each network condition was taken into consideration and translated to state event trigger. The finite state being exclusive for all input combinations means that the design is consistent, the network can be in only one state at a time. This means that this design can be implemented in real-time without causing inconsistency in a network as outlined in the next section.

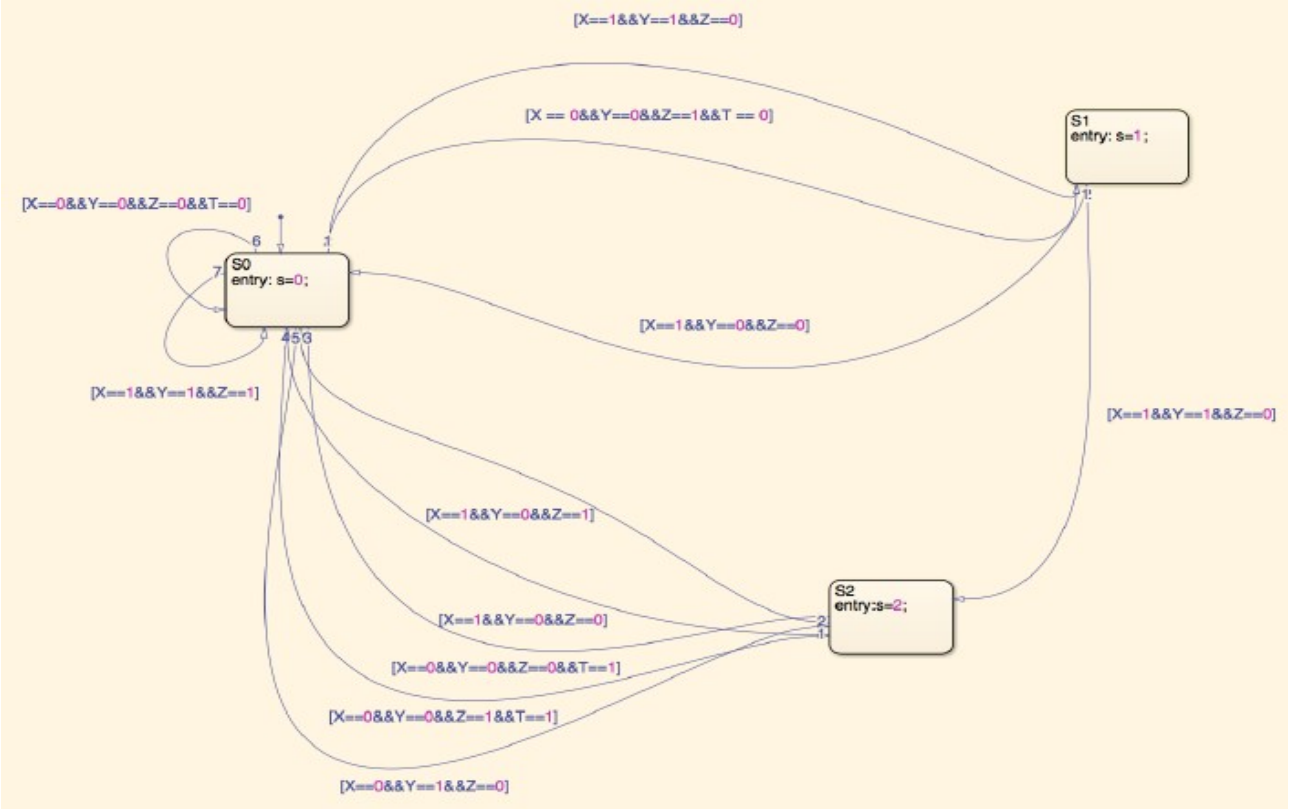


Figure 4: State Transition Model

V. Conclusion and Future Work

This paper proposed a scheme for autonomic management and orchestration of Fog computing networks using finite state machine. The scheme leverages on the work that has been done in integrating OpenFog architecture and ETSI NFV MANO framework to address management and orchestration challenges in Fog computing, and also adds the unique feature of finite state machines for network state event triggers.

We believe the proposed scheme is an important alternative to machine learning that has been popular in networking to address management and orchestration challenges, this design is an option to provide decent intelligence to a network without going through the different stages required to setup a machine learning function to support network decision making.

The finite state machine has been successfully used by the most widely used communication protocols. The next steps in this work is to implement the proposed autonomous edge network management scheme. The finite state machine will be integrated with an ETSI NFV MANO framework in a testbed to test and demonstrate the proposed solution in a real-world environment.

References

[1] Cellular Networks for Massive IoT: Enabling Low Power Wide Area Applications, Stockholm, Sweden:, pp. 1-13, 2016. Availabole at: https://www.gsma.com/membership/wp-content/uploads/2016/03/wp_iot.pdf

[2] Network Functions Virtualisation (NFV) Management and Orchestration, 2014 [Online] Available: http://www.etsi.org/deliver/etsi_gs/NFV-MAN/001_099/001/01.01.01_60/gs_NFV-MAN001v010101p.pdf.

[3] Y.C. Hu, Milan Patel, Dario Sabella, Nurit Spretcher and Valerie Young. Mobile Edge Computing A Key Technology towards 5G, 2015 [Online] Available: http://www.etsi.org/images/files/ETSIWhite-Papers/etsi_wp11_mec_a_key_technology_towards_5g.pdf.

[4] P1934/D2.0, Apr 2018 - IEEE Approved Draft Standard for Adoption of OpenFog Reference Architecture for Fog Computing, 2018

[5] M. Yannuzzi. et al. "Toward a Converged OpenFog and ETSI MANO Architecture" in proc. IEEE Fog World Congress (FWC), 2017

[6] Mowei Wang, Yong Cui, Xin Wang, Shihan Xiao, and Junchen Jiang. "Machine Learning for Networking: Workflow, Advances and Opportunities", IEEE Network, April, 2018.

[7] Frank van Lingen, et al. "The Unavoidable Convergence of NFV, 5G, and Fog: A model driven Approach to Bridge Cloud and Edge" in IEEE Communications Magazine, pp 28 – 35, August 2017

[8] José Santos, Tim Wauters, Bruno Volckaert and Filip De Turck. "Fog Computing: Enabling the Management and Orchestration of Smart City Applications in 5G Networks" in MDPI Entropy Available Online: www.mdpi.com/1099-4300/20/1/4/pdf

[9] Mathias Santos de Brito et. al. “A service orchestration architecture for Fog-enabled infrastructures.” in IEEE International Conference on Fog and Mobile Edge Computing (FMEC), 2017

[10] Lei Xu et. al. “CogNet: A network management architecture featuring cognitive capabilities” in proc. IEEE European Conference on Networks and Communications (EuCNC), 2016.

[11] Wei Jiang; Mathias Strufe; Hans D. Schotten. “Intelligent network management for 5G systems: The SELFNET approach” in proc. IEEE European Conference on Networks and Communications (EuCNC), 2017.

[12] F. N. Koumboulis, D. G. Fragkoulis and L. E. Gikas. “Supervisory control for hierarchical switching of communication protocols” in proc. IEEE South Eastern European Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM), 2017.

[13] Goutham Mallapranda, Ishanu Chattopadhayay and Asok Ray. “Autonomous Navigation of Mobile Robots Using Optimal Control of Finite State Automata” in proc. IEEE Conference on Decision & Control, 2006

[14] Thomas Magedanz. “On the ICT Impacts on Telecommunications - Towards software based networks (SDN/5G)”. Lecture University of Cape Town, 2018.

[15] Pedro Garcia Lopez et. al. “Edge-centric Computing: Vision and Challenges”, ACM SIGCOMM Computer Communication Review, Volume 45, Number 5, October 2015

[16] Arslan Munir, Prasanna Kansakar and Samee U. Khan. “IFCIoT: Integrated Fog Cloud IoT Architectural Paradigm for Future Internet of Things”, IEEE Consumer Electronics Magazine 2017

[17] Wei Yu et. al. “A Survey on the Edge Computing for the Internet of Things”, IEEE Access, Volume: 6, 2017.

Performance Analysis of a Data Communication System Employing Bluetooth and Wi-Fi Standards

Mussawir Ahmad Hosany
Department of Electrical and Electronic Engineering, University of Mauritius
Reduit, Mauritius
m.hosany@uom.ac.mu

Luxshmana Chooramun
Department of Electrical and Electronic Engineering, University of Mauritius
Reduit, Mauritius
luxshmana.chooramun@umail.uom.ac.mu

***Abstract*—Bluetooth and IEEE 802.11b (Wi-Fi) standards both coexist and share the same 2.4GHz frequency range of the unlicensed Industrial Scientific and Medical (ISM) radio band. This coexistence of Bluetooth and IEEE 802.11b can cause interference issues and may result in significant performance degradation if parameters of both standards are not properly set. In this paper, we propose to design a data communication system with standard parameters to allow coexistence of both technologies. Also, the performance analysis of a Bluetooth voice link and Wi-Fi data communication system are carried out with the aid of simulations from MATLAB Simulink®. An audio wav file is used as a source and it is transmitted using Bluetooth voice packets. The performance of the system is observed with Frequency Hopping Spread Spectrum (FHSS) multiplexing techniques. Additive White Gaussian Noise (AWGN) channel is also used in the system and the effect of AWGN is also analyzed during the transmission process. Two different voice packets are used namely HV1 and HV3 voice packets. It is shown that using Bluetooth Standard HV1 voice packet provides better error performance as compared to HV3 voice packet because HV1 packets use 1/3 Forward Error Correction (FEC) while HV3 do not use any FEC. The bit error rate is reduced when Es/No is increased providing improvement in the error performance.**

Keywords—Data Communication, Bluetooth, IEEE 802.11, Wireless LAN, error performance, communication channels

I. INTRODUCTION

Over the past decades in the communication industry wireless networks and systems have been swiftly developing and several experiments were carried out with the transmission of wireless signals. Each type of wireless network operates on a specific set of radio frequencies [1] Two most popular wireless networks are Wireless Local Area Network (WLAN) or Wi-Fi and Bluetooth.

The IEEE 802.11 standard (WLAN) specified an operating frequency of 2.4GHz (ISM) band with maximum data transmission rates of 1 and 2 Mbps. Frequency Hopping Spread Spectrum (FHSS), Direct Sequence Spread Spectrum (DSSS) and Infrared (IR) were defined by the physical layer specifications [2]. In 1999, the IEEE 802.11b standard was launched also known as Wi-Fi and the standard uses DSSS technique where data is transmitted at a rate of up to 11 Mbps using a bandwidth of 22 MHz.

The IEEE 802.15 standard also used the 2.4GHz ISM band. Bluetooth technology employs the WPAN and operates from a frequency range 2.4GHz to 2.4835GHz. The IEEE 802.15.1 standard is also known as Bluetooth and specifies the operation and architecture of Bluetooth devices. The IEEE 802.15.1 WPAN standard is a set of the physical and the MAC layer protocols. The same OSI protocol stack is used for the case of IEEE 802.11 WLANs [2,3].

Bluetooth employs Frequency Hopping Spread Spectrum (FHSS) and Gaussian Frequency Shift Keying (GFSK) modulation techniques. The channel is divided into 625μs slots using Time Division Multiplexing (TDM) technique and the system will hop between different slots [4]. Spread spectrum techniques are adopted to reduce the interference. DSSS technique is used for Wi-Fi whereas FHSS technique is used for Bluetooth [5].

The aim of this research work is to analyze the coexistence of a Bluetooth signal and a Wi-Fi signal transmitted with the same unlicensed 2.4GHz (ISM) frequency band when a Bluetooth transmitter also known as the Master device transmits a voice signal using SCO links to the receiver known as the Slave device. Three different types of SCO voice packets are considered in this project. The performance analysis of IEEE 802.11b (Wi-Fi) and Bluetooth version 1.0 are carried out with FHSS modulation techniques. Forward Error Correction (FEC) technique is employed for error control in the transmission process. The Bit Error Rate (BER) is calculated and compared to the Additive White Gaussian Noise (AWGN). Matlab Simulink® is used as a platform to simulate the model.

Section II briefly introduces the IEEE 802.15 Bluetooth standard and section III discusses the various coding and modulation schemes employed in both standards. Section IV presents our software design process and implementation of our data communication system. Section V presents the simulated results of our proposed system and section VI concludes the paper.

II. BLUETOOTH STANDARD

Bluetooth has come to the fore as a wireless communication technology for successfully bringing the interconnection between various computer peripherals and has become a useful asset for mobile phones, computers and Personal digital assistant (PDA). Bluetooth is a standard for short distance wireless communication. It supports a range of 10 meters to communicate [4].

A piconet in the context of Bluetooth is a very important term where two or more devices are synchronized to the similar hopping sequence, where one device functions as a master and the others as slaves. A piconet is formed when two or more devices use the same channel. There are a maximum of eight active devices representing a 3-bit address which can be employed in Bluetooth.

Point-to-point fashion and point-to-multipoint fashion are used where the piconets slaves only communicate with the

piconet master. The frequency sequence hops and clocks of the slaves must synchronize with the master when connected.

A scatternet is formed when multiple piconets are connected. A Bluetooth device can take part in multiple piconets at the same instant and one piconet can also act as a part of another piconet. Moreover, a device may act as both a master and a slave within the scatternet.

The Bluetooth protocol stack consists of several layers namely Radio, Baseband, Link manager protocol, Service discovery and RFCOMM. In the Bluetooth technology the physical link consists of two types. Bluetooth allows voice and data transmissions. For Bluetooth voice transmission a synchronous connection-oriented (SCO) link is used and for data transmission asynchronous connection-less (ACL) link is used [5].

A SCO link is a point-to-point connection where reserved time slots called as T_{SCO} time slots are used by the master transmitter at regular interval to transmit packet to the slave in a piconet. The three types of voice packets used are HV1, HV2 and HV3. The voice packets transmit only every sixth slot at a time period of 3.75 ms. Forward Error Correction (FEC) is used by HV1 and HV2 voice packets at a rate of 1/3 and 2/3 respectively, whereas ACL link use 2/3 rate only. HV3 uses no FEC mechanisms. 10 bytes is used by HV1 packet. 20 bytes is used by HV2 packet and 30 bytes is used by HV3 packet to carry information. There is no retransmission of SCO voice packets instead Continuous Variable Slope Delta (CVSD) encoding speech is used which a very complex one. As per Bluetooth standard 100% of the bandwidth is used by HV1 packets, 50% is used by HV2 packets and 33.3% is used by HV3 packets [5,6]. Table 1 gives the specifications for the various Bluetooth voice packets.

TABLE 1: BLUETOOTH VOICE PACKET SPECIFICATIONS

Type	Payload (Bytes)	FEC	CRC	Symmetric Maximum Rate (kbps)
HV1	10	1/3	No	64.0
HV2	20	2/3	No	64.0
HV3	30	No	No	64.0
DV	10 + (0-9)D	2/3D	Yes D	64.0 + 57.6 D

III. CHANNEL CODING AND MODULATION SCHEMES IN BLUETOOTH AND WI-FI STANDRDS

Channel coding or Forward Error Correction (FEC) schemes increase the noise immunity of Bluetooth. A receiver performs error correction with FEC by adding redundant information to the data stream during transmission. Three types of FEC are used in Bluetooth technology namely 1/3 rate FEC, 2/3 rate FEC and Automatic repeat request (ARQ) which are applied to HV1 data packets.

Bluetooth technology uses Continuous Variable Slope Delta (CVSD) speech coding technique and 64 kbps PCM format (A-law or μ-law). Modulation schemes are used so as to increase the transmitted signal's immunity to radio channel noise and other interference. In Bluetooth, Gaussian Frequency Shift Keying (GFSK) is employed. For more details the reader is referred to [7,8]. Spread spectrum modulation is another type of modulation employed by the IEEE 802.11 Wi-Fi and Bluetooth standards [9,10]. Some advantages of spread spectrum are listed as follows:

- It has the capability to protect against multipath interference.
- Excellent anti-interference characteristics.
- Privacy
- Ability to overlay a spread spectrum system over a frequency band which are in use.

There are two well-known spread spectrum schemes known as Frequency Hopping Spread Spectrum (FHSS) employed by Bluetooth standard and Direct Sequence Spread Spectrum (DSSS) employed by IEEE 802.IIb (Wi-Fi) [1]. FHSS is a spread spectrum technique where the center frequency of transmission changes periodically according to the pseudo-random code sequence. In Bluetooth transmission, the signal hops 1600 per second among the 79 channels equally spaced with 1 MHz and with a time slot of 625μs.

The Bluetooth Special Interest Group (SIG) introduced FHSS, shown in Fig. 1, where the objective is to change the frequency periodically when the channel frequency is shared by each member of a piconet according to the following formula

$$f = 2402 + k\ MHz, \qquad k = 0, \ldots, 78.$$

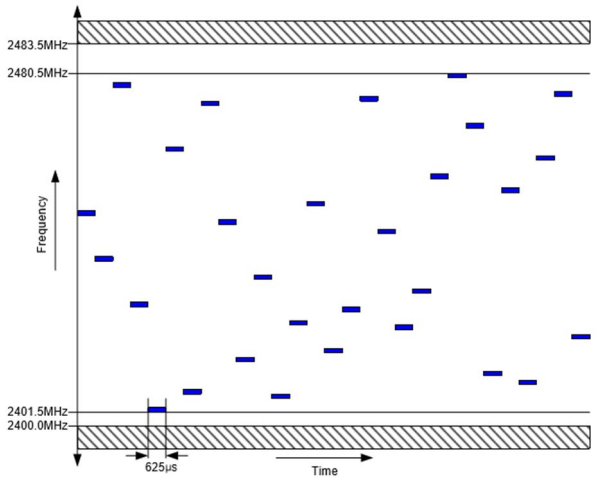


Fig. 1. FHSS in Bluetooth

FHSS takes place according to the pseudo-random code sequence [11]. A pseudo-random sequence depends on the Bluetooth Master transmitter clock and the Slave receiver MAC address and produces the same pseudo-random sequence. To select the sequence of frequencies, these pseudo-random binary codes are used, the transmitter then use these codes for frequency hopping. The pseudo-random numbers are obtained by an initial number which can be considered as the seed. Pseudo-random codes need to be generated in order for a receiver to decode the transmission.

DSSS is another kind of spread spectrum modulation used by IEEE 802.11b (Wi-Fi)where 22MHz of bandwidth is used and data is transmitted at a rate of up to 11Mbps. IEEE 802.11b standard keeps the same frequency utility over time

and can use any of the 22MHz wide sub channels in the range of the 83.5MHz available [12,13].

IEEE802.11b makes use of bit sequence known as chipping code to spread along the frequency band. It uses Barker coding and Complementary Code Keying (CCK). The DSSS systems provide a wireless LAN with both 1 and 2 Mbps data rate and modulation scheme of either differential binary phase shift keying (DBPSK) or differential quadrature phase shift keying (DQPSK) [14].

For higher rate physical layer IEEE 802.11b employs complex modulation schemes giving data rates of 5.5 Mbps and 11 Mbps. 8-chip CCK modulation scheme is used to provide these high data rates. To achieve a data rate of 5.5Mbps, 4 data bits per symbol are able to transmit by the use of 1.375 MSps and to achieve 11Mbps data rate, 8 data bits per symbol are transmitted also at 1.375 MSps rate [14].

IV. Software Design and Implementation

The software design of our proposed data communication system is composed of several processes discussed in the above sections and these processes are used to design and implement the model on MATLAB Simulink® with the DSP and Communication system toolboxes [15]. SCO packets are used for the voice transmission**.**

The Simulink design is shown in Fig. 2. The system consists of a master transmitter, AWGN channel, 802.11b as interferer, CSVD speech coding, GFSK modulation, Slave receiver, BER meters, time scope, spectrum analyzer and spectrogram. The system is built by a master transmitter that uses an input file which is in WAV format at 8000 Hz sampling rate. A time scope is used to display the input WAV file. The output signal is fed to an FIR filter and then speech encoded using CVSD.

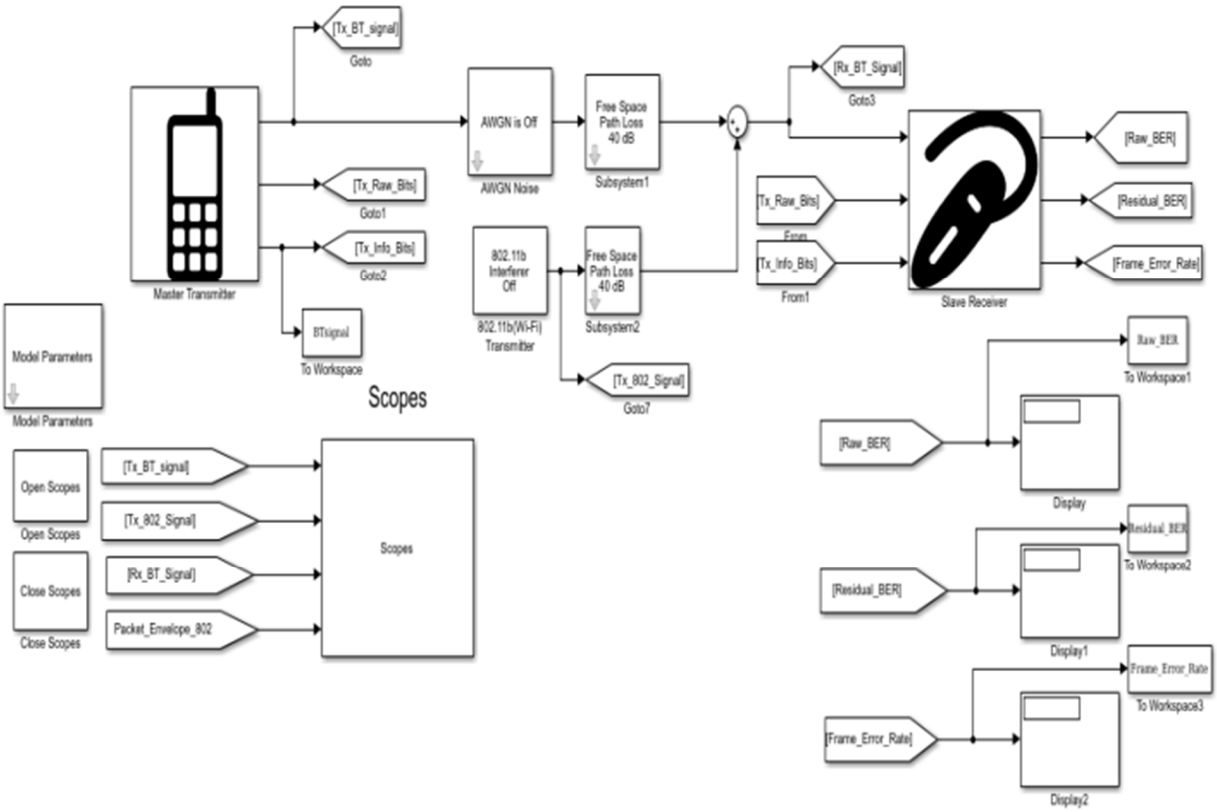


Fig. 2. Data communication system employing Bluetooth and Wi-Fi standards in Matlab Simukink®

The generated voice packets called HV1, HV2 and HV3 are then passed to the FEC blocks for error correction after travelling over the AWGN channels. HV1 and HV2 uses FEC with code rates 1/3 and 2/3 cyclic codes respectively whilst HV3 do not have FEC as per the Bluetooth standard.

The hopping sequence generator block is a masked subsystem which performs frequency hopping techniques. It uses pseudo-random hopping sequence techniques through the 79 channels as specified in the Bluetooth standard [16-18]. To generate the GFSK signal a Continuous phase modulation (CPM) block is used from the communication system toolbox to set the modulation index parameter to 0.32 and specifies 100 sample per symbol and these samples produce 62500 samples by simulation which results in slot rate of 625μs as per the Bluetooth standard. The baseband signal is multiplied by one of the 79 possible carriers to avoid interference with other devices operating on the same ISM band and the carriers have frequencies in the range of $\pm$ 39 MHz. The frequency pulse shape parameter is set to Gaussian model.

The AWGN block is used to add additive white Gaussian noise (AWGN) to the system with symbol noise energy (Es/N0) and bit to noise energy (Eb/N0). The noise energy is a mask subsystem so as to either enable noise or disable noise. The channel has a bandwidth of 100MHz which results in increase in simulation time.

The 802.11b uses a bandwidth of 22MHz due to direct sequence chip rate [19]. The 802.11b Wi-Fi Interferer is a masked subsystem which generates 22MHz. Wi-Fi is packet based which uses the Poisson process. The function of the slave receiver is to retrieve the radio signal that is transmitted and is more complex as compared to the transmitter since there is a need for synchronization and detection of errors. The slave receiver performs the same operations as the master transmitter but in the reverse order. The model is assumed to be completely synchronized.

The subsystem that performs GFSK demodulation allows generation of 366 bits once every 6 slots. Following demodulation there is extraction of the access code, header and payload. The 8-bit HEC is revealed by the header which is de-repeated and used for checking the 10-bit header info.

Due to channel impairments a frame is accepted if HEC does not succeed in matching the header info or in the access code there is correction of less than 57-bits. Following is the CVSD speech decoder processing and Bit Error Rate (BER) calculation [20]. The Scopes subsystem display the output of the model after the simulation has run. It consists of a time scope and a spectrum analyzer. The "to multimedia file" block writes audio samples to a multimedia file. It is the output wav file and is saved after the MATLAB simulation is run.

V. Simulation Results and Discussions

The proposed model of Fig. 2 is run to provide experimental results. The unpredictability between the bits in the transmitted signal and the receiver signal is displayed by the Raw bit error rate. The spectrum analyzer and time scopes are used to display the results where the spectrum displays the transmitted Bluetooth signal which is narrowband, together with the IEEE 802.11b signal. A dynamic plot of packet frequency over time is displayed using the spectrogram. The simulation can be run several times to obtain different results when parameters are changed. The results consist of

- A performance analysis of Bit error rate (RawBER) against Es/No (symbol energy to noise energy) where Es/No is varied, keeping *AWGN on* and *IEEE 802.11b off* using HV1. Results also include Timing diagrams and spectrograms.

- The effect of IEEE 802.11b during Bluetooth voice transmission including timing diagrams, channel spectrogram and spectrum.

The Error Rate Calculation block from the communication system toolbox is used to calculate the bit error rate, where the input data from the transmitter and receiver are compared.

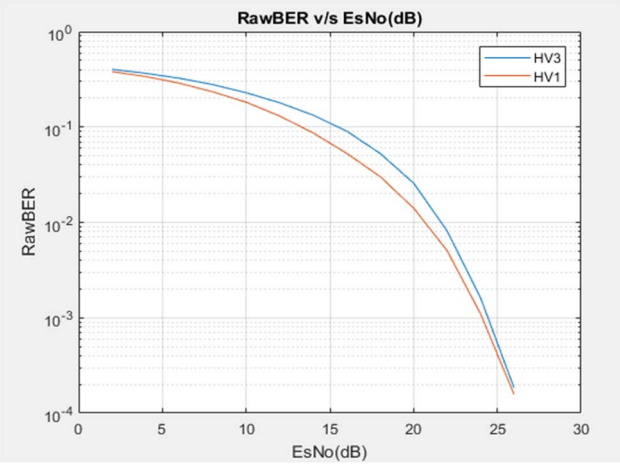


Fig. 3. RawBER against Es/No for HV1 and HV3

From Fig. 3 it can be seen that HV1 packet has lower Es/No compare to HV3 packet and both have a "waterfall-like" shape. At RawBER of $6x10^{-2}$, it is observed that Es/No value for HV1 packet is 15dB and for HV3 packet is about 17dB which is a difference of 2dB. This means that HV1 packet has a better performance and a higher error protection compare to HV3 packet.

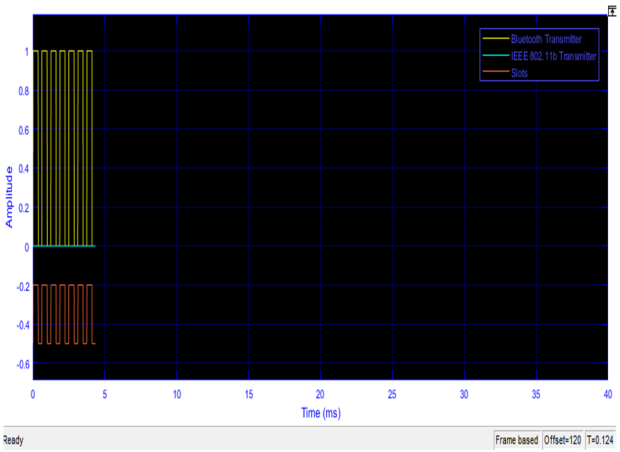


Fig. 4. Timing diagram for HV 1 packets

With the transmission of HV1 packets the value of Es/No is kept at 2 dB and during the simulation run the timing diagram of Fig. 4 was recorded. The timing diagram shows the slots of Bluetooth and IEEE 802.11b and it can be observed that the slot for IEEE 802.11b has zero amplitude because IEEE 802.11b was kept off. The timing diagram offset was set to 120.

The spectrum analyzer block is taken from the DSP system toolbox for displaying the frequency spectra of signals and it can be configured by changing its settings. It is also used to display the spectrogram. Fig. 5 shows that at Es/No = 26dB the channel spectrogram yields clear Bluetooth voice packet transmissions for HV1 packets over the bandwidth of ±50 MHz.

The Bluetooth transmission lines for HV3 packets can be clearly seen in Fig. 6 for a high value of Es/No = 26dB. This is due to the high signal energy that was used resulting in fewer errors during transmissions.

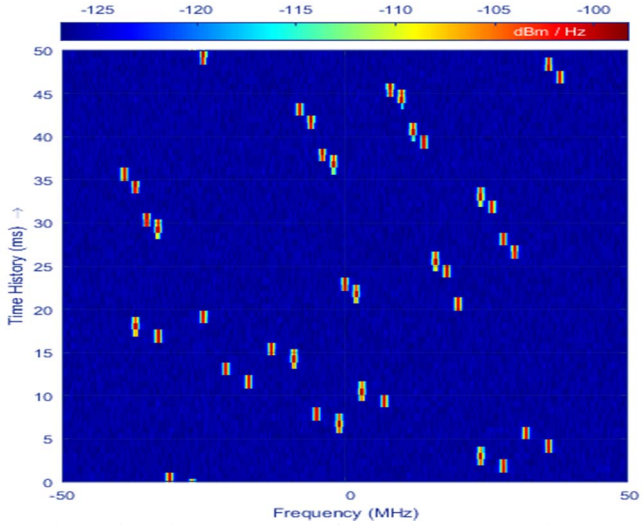


Fig. 5. Channel spectrogram at Es/No=26 dB for HV 1 packets

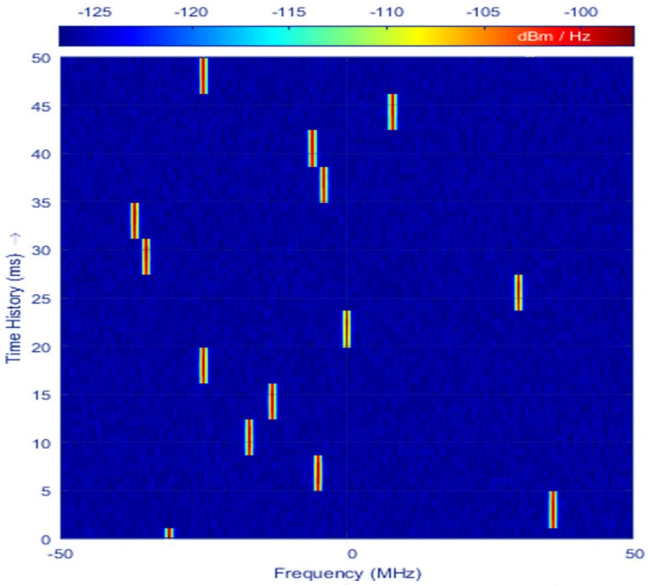


Fig. 6. Channel spectrogram at Es/No=26 dB for HV 3 packets

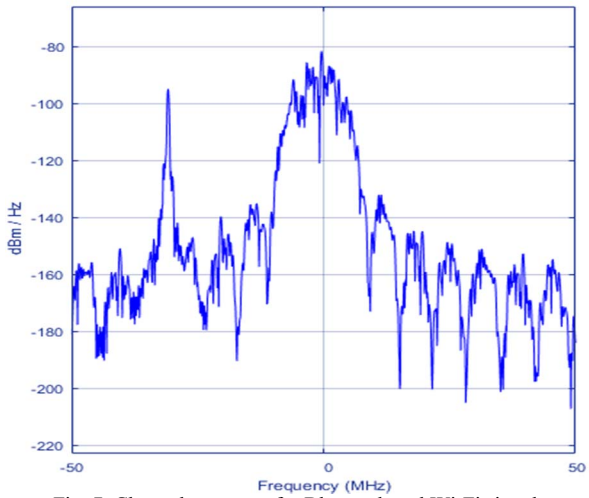


Fig. 7. Channel spectrum for Bluetooth and Wi-Fi signals

Fig. 7 shows the channel spectrum for both the Bluetooth and IEEE 802.11b signals. As IEEE 802.11b (Wi-Fi) uses a bandwidth of 22MHz, it coexists with the Bluetooth signals when the time slots of both signals are the same. The highest single peak is the Bluetooth signal. The simulation was performed with frequency number set to 38 and the packet rate set to 200.

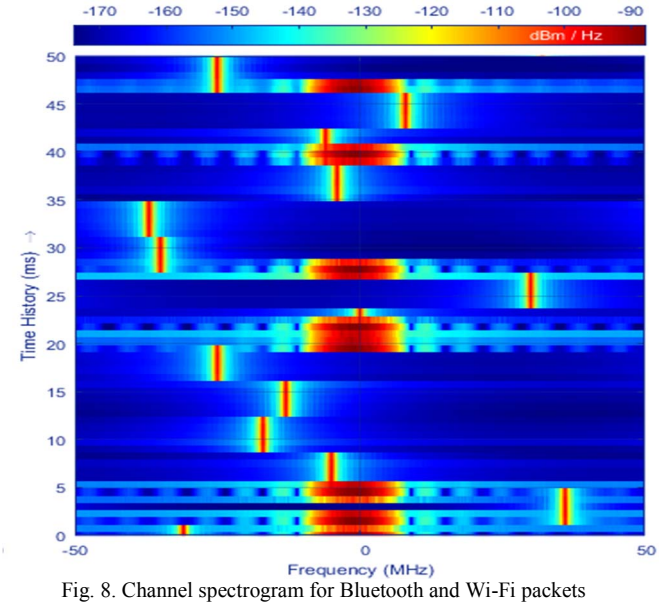


Fig. 8. Channel spectrogram for Bluetooth and Wi-Fi packets

In the channel spectrogram of Fig. 8 the thin vertical lines indicate the Bluetooth voice transmissions while for IEEE 802.11b the packets are thicker ones. It can be seen that both channel packets coexist without interference during transmission.

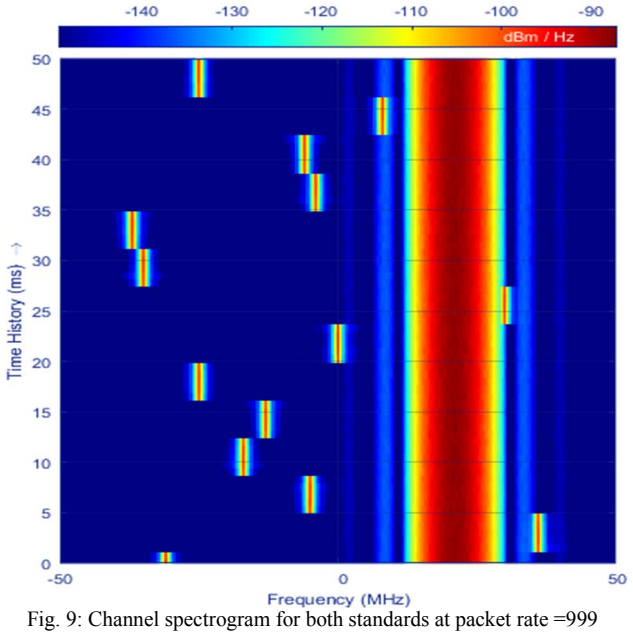


Fig. 9: Channel spectrogram for both standards at packet rate =999

It can be observed from Fig. 9 that the IEEE 802.11b is permanently on during the transmission as the thick blocks are continuous. Both Bluetooth and Wi-Fi packets are coexisting during the transmission as expected. The long thick blocks are the IEEE 802.11b slots, with a packet rate of 999 and the Bluetooth packets are the thin lines scattered over the entire bandwidth.

VI. Conclusions

In this paper the performance of a data communication system employing Bluetooth and Wi-Fi standards has been analyzed. We have designed the system using in-built Matlab Simulink® blocks as per the standards. The model is half-duplex which means that the master device can only transmit and the slave device can only receive the signal. The model also makes use of AWGN channel. Simulation results have been presented and it was observed that HV1 voice packets provide better error performance as compared to HV3 ones because HV1 packets use 1/3 FEC while HV3 ones do not use any FEC. The bit error rate is reduced when Es/No is increased resulting in improvement in the error performance. We have also shown that both Bluetooth and Wi-Fi packets can coexist, without interference, by properly setting the model parameters such as timing offset, Es/No, packet data rate and bandwidth. This designed Simulink model can find applications in the testing process of Bluetooth and Wi-Fi standards.

References

[1] A. Goldsmith, Sample Chapters from "Wireless Communications", Cambridge University Press, 2005 [ONLINE] Available at: https://wsl.stanford.edu/~andrea/Wireless/SampleChapters.pdf

[2] E. Ferro, F. Potorti, "Bluetooth and Wi-Fi wireless protocols: a survey and a comparison.", IEEE Wireless Communications magazines, Vol. 12, Issue 1, pp 12-26, March 2005.

[3] K. Guntass, "Bluetooth and Wi-Fi Interference: Simulations and Solutions", International Journal of Advanced Research in Computer and Communication Engineering, Vol.3, issue 9, pp 8024-8027, September 2014.

[4] Silan Liu, "Bluetooth Technology" [ONLINE] Available at: http://progtutorials.tripod.com/Bluetooth_Technology.htm#_Toc41989851

[5] Vinod Vokkarane, "Architecture and Protocols Used in Common Bluetooth Systems" [ONLINE] Available at https://pdfs.semanticscholar.org/7ca7/9045b835a2504304e995a47dcce3de85ed26.pdf

[6] J. Schiller, "Mobile Communications", Second edition, Pearson edition limited, ISBN: 978-0-321-12381-7, USA, 2003.

[7] S. Ali Alghamdi, N.Ali Mahdi, A.Zohdy Mohamed, "Robust non-coherent Demodulation scheme for Bluetooth voice transmission using linear, Extended, and unscented kalman filtering, "Journal of signal and information processing, Scientific Research Publishing, Vol. 6, pp. 9-27, 2015.

[8] H. S. Gerez "Implementation of Digital Signal Processing: Some Background on GFSK Modulation" Version 5, 9 March 2016.

[9] S. S. S. Sugi,C.V. Joe, "Spread Spectrum Modulation Techniques using MATLAB", International Journal of Innovative Research in Technology, Vol 2, Issue 3, pp 47-52, Aug 2015.

[10] A. Kurawar, A. Koul, V. T. Patil, "Survey of Bluetooth and Applications", International Journal of Advanced Research in Computer Engineering & Technology. Vol. **3**, pp 2832–2837, August 2014.

[11] mobiHoc, "A Pseudo Random Coordinated Scheduling Algorithm for Bluetooth Scatternets" [ONLINE] Available at: http://slideplayer.com/slide/9484556/

[12] N. Golmie, R.E. Dyck, A. Soltanian, O. Rébala,"Interference Evaluation of Bluetooth and IEEE 802.11b Systems", Journal of Wireless Networks, Vol. 9, Issue 3, pp 201-211, Springer-Verlag, 2003.

[13] M. Abusubaih, M. Ayyash, "Mutual interference between bluetooth and 802.11n MIMO devices", 2012 19th International Conference on Telecommunications ICT 2012*, no. Ict, 2012.*

[14] R. J. Punnoose, R. S. Tseng, and D. Stancil, "Experimental results for interference between Bluetooth and IEEE 802.11b DSSS systems," Proc. IEEE VTC , Atlantic City, NJ, pp. 67–71, Oct 2001.

[15] D. K. Chaturvedi, "*Modelling and Simulation of Systems Using MATLAB and Simulink*", CRC press, pp 621-665 ISBN: 978-1-4398-067-2, 2010.

[16] Anil Mathew, Nithin Chandrababu, Khaled Elleithy, Syed Rizvi. Interference of 802.11b WLAN and Bluetooth: Analysis and Performance evaluation, *"International Journal of Computer Networks & Communications (IJCNC),* Vol.2, No.3, pp 140-149, May 2010.

[17] Zahir Aalam and B.K Mishra, " Simulation and Throughput Analysis of Multiple IEEE 802.15.1 Devices in Presence of Interference", International Journal of Computer Applications, pp 48-52, 2013.

[18] Zahir Aalam, and B. K. Mishra, "IEEE 802.15.1 Simulation and BER Analysis under the interference", IJCA Proc. on Int. Conf. and Work. on Emerg. Trends in Tech. (ICWET), pp. 24-28, 2011.

[19] Implementing wireless LAN using IEEE 802.11b standard. [ONLINE] Available at

http://www.ece.mtu.edu/ee/faculty/rezaz/wlps/papers/temp/packetstructure.pdf

[20] BER Bit Error Rate Tutorial and Definition, RF Technology & Design. [ONLINE] Available at:

http://www.radio-electronics.com/info/rf-technology-design/ber/bit-error-rate-tutorial-definition.php

Retinal Vessel Tortuosity Characterisation Using Boosted-Linear Discriminant Analysis

Temitope Mapayi
Department of Electrical Engineering
Mangosuthu University of Technology
Umlazi, Durban, South Africa.
Email: mapayi.temitope@mut.ac.za

Emmanuel Gbenga Dada
Department of Computer Engineering
University of Maiduguri,
Nigeria
Email: gbengadada@unimaid.edu.ng

Moses Olaifa & Pius Adewale Owolawi
Department of Computer Systems Engineering
Tshwane University of Technology
Pretoria, South Africa
Email: olaifamo@tut.ac.za; owolawipa@tut.ac.za

***Abstract*—Retinal vessel tortuosity is an early indicator of different retinopathies. Although various automated methods in determining retinal vessel tortuosity have been proposed in the literature, there are needs for further study. This study extracted three different features namely distance metric, normalized hybrid metric and non-normalized hybrid metric from the thinned vessels. The weights of vessel data samples were dynamically updated using the Adaboost with linear discriminant analysis (LDA) and the feature correlation was used to facilitate the selection of the best feature combination at each of the boosting iteration rather than a single feature that minimizes the weighted error at each of the iterations. Adaboost with LDA method is then used for the classification of the retinal vessels as either tortuous or normal using a majority voting method. The proposed method achieves the accuracy rate of 100% for the training sample sizes of 70%, 80% and 90%.**

***Keywords*—Characterisation, Directional Changes, Morphological, Retinal Vessel, Supervised Learning, Tortuosity.**

I. Introduction

The use of retinal fundus images in ophthalmology for the diagnosis of retinopathy of prematurity (ROP) and diabetic retinopathy (DR) [1] has gained wide acceptability because it is safe and cost-effective [1].

There are several important anatomic structures in the human retina and their robust detection and characterisation are needful for the proper identification of healthy or diseased retina. One of the anatomic structures are the retinal blood vessels. Different morphological characteristics of the blood vessels in the retinal fundus images such as vessel shape, branching pattern, width, tortuosity and retinal lesions appearance can be associated with the presence of retinopathies and are used for diagnosis to determine the state of the patients as either diseased or healthy [2], [3]. These are also used to monitor the progress of the patients that are undergoing treatment [2], [3]. Tortuosity is identified as a major morphological characteristic of the retinal vessels that serves as an initial indication of DR and ROP [4], [5], even when there are no detectable manifestation of pathologies in the retina [3], [6] (See. Fig. 1).

It was however noted that the process of manually detecting the retinal vessels and the analysis of their features, by the medical experts that are very few, are time consuming and laborious when attending to a large number of patients [6], [7], [8].

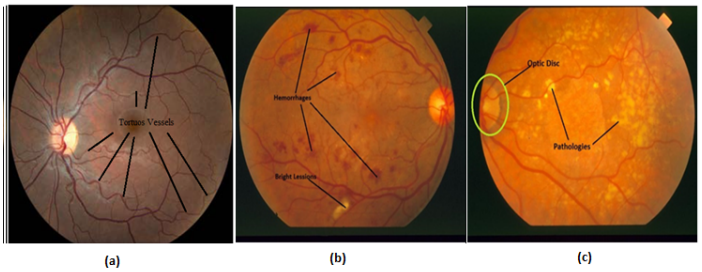


Fig. 1: (a) Retinal Image With Tortuous Vessels (b) Retinal Image With Lesions (c) Non-Proliferated Retinal Image With Pathologies.

It has however been noted that a timely efficient and highly accurate retinal vessel detection approach is required for a reliable vessel morphological feature characterisation [6]. Several automated methods have been implemented in the literature for the detection of the retinal vessels [9], [10], [11], [12], [13], [14] and their tortuosity characterisation [3], [6], [15], [16], [17] (see Fig. 2).

The different automated methods used for determining vessel tortuosity can be divided into unsupervised and supervised approaches. The unsupervised approach consists of methods based on different models such as chain code [18], vessel inflection points [19], second derivatives [20], integral curvature [21] and arc-chord ratio [6], [15], [22], [23], [24], [25]. Other unsupervised approach are hybrid-based technique combining vessel-thickness with curvature of vessel boundary [26], hybrid-based technique combining changing curvature sign and arc to chord ratio Grisan et al. [27] and hybrid-based technique combining distance metric and stationary points in [6].

Different supervised learning methods have also been proposed in the literature [3], [28], [29]. Turior & Uyyanonvara [29] investigated the use nave Bayesian classifier, k-nearest neighbour (kNN) classifier, and K-means clustering algorithm using some features for the tortuosity characterisation of the retinal vessels in fundus images. KNN classifier achieved the highest performance with the average accuracy rate of 87.3%. An improved chain code was used to compute curvature metrics as features in [28]. These features were fed into kNN and nave Bayesian (NB) classifiers for classification purposes [28]. KNN classifier has the highest performance with sensitivity

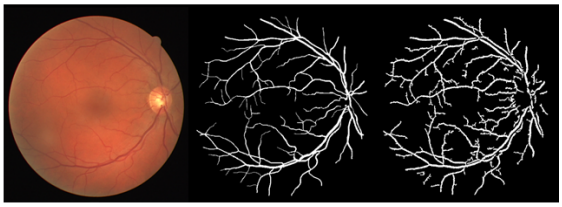


Fig. 2: (a) Colored Fundus Retinal Image (b) Manually Segmented Retinal Image (c) Automated Segmented Vessel in Retinal Image [13]

Fig. 3: (a) Vessel Segment in Colored Fundus Image (b) Binarised Vessel Segment (c) Thinned Vessel Segment

rate of 97.8% and and accuracy rate 93.6%. A comparison of four different classifiers namely kNN, NB, support vector machine (SVM) and decision tree were investigated in [3]. Both kNN and NB classifiers achieved the highest accuracy rate of 100% when their training vessel data-samples are 70%, 80% and 90%.

Even though the aforementioned methods have made considerable progress, their results require further need for improvement. Consequently, this paper presents a study on the determination of retinal vessel tortuosity in fundus images using Adaboost with linear discriminant analysis (LDA).

The remaining part of this paper is organized as follows: The methods and techniques used in this paper are presented in section II. The experimental setup, results and the discussion of the results are presented in section III and section IV presents the conclusion of the study.

II. Methods and Techniques

This section presents the use of the three different features based on geometrical properties of the retinal vessels used in [3] and AdaBoost based on linear discriminant analysis classifier for the classification of retinal vessel as either tortuous or non-tortuous in fundus images.

The automated delineated binarised retinal vessels from the colored fundus images are used as the input images for this study (see Fig. 2). The binarised vessels are thinned (see Fig. 3) using the algorithm proposed in [3], [6], [30] and a moving mean filter is used to remove the smaller swings due to noise in the thinned vessel as implemented in [3], [6]. The three different extracted features computed in [3] and [6], namely distance metric, normalized hybrid metric and non-normalized hybrid metric features are used in this study. The frequency of the retinal vessel directional changes computed using Knuth-Morris-Pratt (KMP) algorithm [31] were combined with the distance metric to compute the normalized hybrid metric and non-normalized hybrid metric features (see [3] & [6] for details). All the three extracted features are fed into the classifier for the classification purpose to determine the retinal vessels as either tortuous or non-tortuous in the fundus images.

Linear discriminant analysis (LDA)is a machine learning technique used for both classification and feature space dimensionality reduction. Adaboost is an ensemble method used to combine a set of weak classifiers to achieve a very good accurate classifier. In this paper, we will be using a boosting method because it has been discovered that it is one of the successful methods for creating ensemble classifiers. Hence, AdaBoost with LDA is used to determine whether the retinal vessels are tortious or not tortious. Given that 'p' is the number of training vessel data-samples, an initial weight W_0(j)=1/p is assigned to each training sample x_j. At the d_{th} iteration, Adaboost technique utilizes the vessel data-samples x_j that minimizes the weighted error E_d according to W_d(j) to determine the classifier h_d to be trained. Given that the ground truth labels are y_j and the normalization factor is Z_d, α_d is computed as

$$\alpha_d = \frac{1}{2}\log\frac{1-E_d}{E_d} \tag{1}$$

and Equation 2 updates the weights as:

$$W_{d+1}(j) = \frac{W_d(j).exp - \alpha_d y_j h_d(x_j)}{Z_d} \tag{2}$$

The different weights are changed accordingly such that the weights of appropriately classified vessel data-samples are increased while the different weights of the inappropriately classified vessel data-samples are decreased. Given that the hypothesis of the t^{th} weak classifier is h_t, the hypothesis of a strong classifier H(x)is computed as:

$$H(x) = sign\left(\sum_{t=1}^{T}\alpha_t h_t(x)\right) \tag{3}$$

Adaboost with LDA is utilized to update the weights of training vessel data-samples dynamically. A nearest class-center classifier, which is the distance between the test vessel data and the centers of the training vessel data of each class, is utilized for the computation of the training error and test error. Since we are using weighted LDA, the weights are taken into consideration while combining it with Adaboost, and utilizes the feature correlation to facilitate the selection of the best feature combination at each of the boosting iteration rather than a single feature that minimizes the weighted error at each of the iterations. The scatter matrices are then computed. The the intra-classes (i.e. within class) scatter matrix is computed as:

$$Class_{wt} = \sum_{c=1}^{d}\sum_{i\epsilon S_c}^{d} P(j)(x_j-\mu_c)(x_j-\mu_c)^T \tag{4}$$

and the inter-classes (i.e. between classes) scatter matrix is computed as:

TABLE I: Performance Comparison of the AdaBoost with Different Classifiers in the Literature Over Different Training Sample Sizes

Method	Accuracy Rates (%)								
	Training Samples (%)	10	20	30	40	50	60	70	80	90
SVM [3]		66.67%	37.5%	71.42%	73.33%	76%	75%	86.67%	87.5%	83.3%
K-Nearest Neighbour [3]		82.2%	82.5%	82.9%	86.67%	92%	95%	100%	100%	100%
Naive Bayesian [3]		55.8%	72.5%	82.9%	86.67%	96%	95%	100%	100%	100%
Decision Tree [3]		82.2%	82.5%	82.35%	85.7%	72%	85%	86.7%	80%	83.3%
Proposed Method: AdaBoost With LDA		60%	67.5%	80%	86.67%	92%	85%	100%	100%	100%

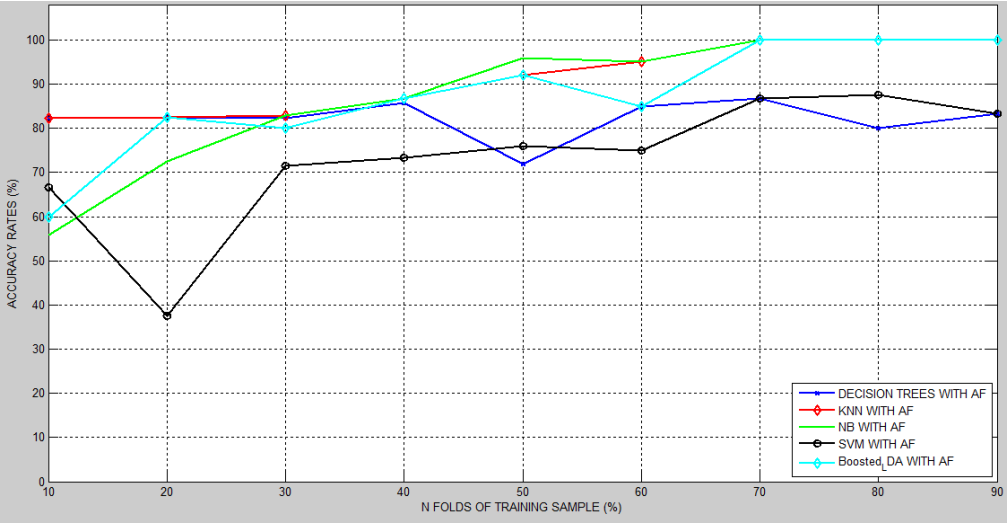


Fig. 4: Comparison of the Accuracy Rates of AdaBoost with LDA And Some Classifiers Over Different Training Sample Sizes

$$Class_{bt} = \sum_{c=1}^{d} \left(\sum_{j \epsilon S_c}^{d} P(j) \right) (\mu_c - \bar{x})(\mu_c - \bar{x})^T \quad (5)$$

where P(j) is the weight assigned to each j_{th} training sample and the weighted class mean μ_c is computed as:

$$\mu_c = \frac{1}{\sum_{j \epsilon S_c}^{d} P(j)} \sum_{j \epsilon S_c}^{d} P(j) x_j \mu \quad (6)$$

and $\bar{x}$ is computed as:

$$\bar{x} = \sum_{j \epsilon S_c}^{d} P(j) x_j \mu \quad (7)$$

A majority voting method is then used for the final decision rule to determine the class of the retinal vessel as either tortuous or normal.

III. Experimental Results and Discussion

The experiments conducted in this study were performed using a MATLAB 2014a on an Intel Core i5 2410M CPU, 2.30 GHz, 4GB of RAM. The dataset used in the study is made up of 50 digital fundus images containing real vessel segments each. The expert's ground truth of the dataset are provided by a consultant ophthalmologist. The dataset are divided into training set and testing set as described in Table I. The classifier was trained using the training set and the testing set were used in the testing phase for classification purpose. The performance metric used in this study is accuracy measure and this is shown in Equation (8)as:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (8)$$

where TP = True Positive, TN = True Negative, FP = False Positive and FN = False Negative.

The result presented in Table I shows the performance comparison of the proposed method in this paper with some

other supervised learning methods implemented on the same dataset in [3]. SVM classifier achieved lower accuracy rates ranging from 37.5% to 83.3% when compared with the the method proposed in this paper that achieved higher accuracy rates ranging from 67.5% to 100%. SVM however performed better than the proposed method when the training sample size of 10% (see Table I & Fig. 4). Decision tree classifier achieved lower accuracy rates ranging from 72% to 85.7% when the training sample size is between 40% and 90% and the method proposed in this paper achieved higher accuracy rates ranging from 85% to 100%. Decision tree classifier however achieved higher accuracy rates when the training sample size is between 10% and 30% when compared with the the method proposed in this paper. KNN classifier, NB classifier and the method proposed in this paper achieved the highest accuracy rates of 100% when the training sample size range between 70% and 90%.

IV. Conclusion

This paper presented a boosted linear discriminant analysis method for the classification of retinal vessels as either tortuous or normal. Three different features based on geometrical properties were extracted from the retinal vessels and fed into Adaboost classifier with linear discriminant analysis. The method proposed in this paper achieved the highest accuracy rate of 100% for the training sample sizes of 70%, 80% and 90%. This study also showed that the use of boosted linear discriminant analysis for retinal vessel tortuosity characterisation achieved very promising results. Our future work will investigate the use of other morphological features to characterise retinal vessels in retinopathy patients.

Acknowledgement

We would like to appreciate the consultant ophthalmologist who provided the professional assistance in creating the ground truth of the dataset used in this study

References

[1] M. D. Abràmoff, M. K. Garvin, and M. Sonka, "Retinal imaging and image analysis," *Biomedical Engineering, IEEE Reviews in*, vol. 3, pp. 169–208, 2010.

[2] V. S. Joshi, "Analysis of retinal vessel networks using quantitative descriptors of vascular morphology," 2012.

[3] T. Mapayi and J.-R. Tapamo, "Performance comparison of supervised learning methods for retinal vessel tortuosity characterisation," in *Information Communications Technology and Society (ICTAS), 2018 Conference on*. IEEE, 2018, pp. 1–7.

[4] A. Chakravarty and J. Sivaswamy, "A novel approach for quantification of retinal vessel tortuosity using quadratic polynomial decomposition," in *Medical Informatics and Telemedicine (ICMIT), 2013 Indian Conference on*. IEEE, 2013, pp. 7–12.

[5] B. V. Davitt and D. K. Wallace, "Plus disease," *Survey of ophthalmology*, vol. 54, no. 6, pp. 663–670, 2009.

[6] T. Mapayi, J.-R. Tapamo, S. Viriri, and A. Adio, "Automatic retinal vessel detection and tortuosity measurement," *Image Analysis & Stereology*, vol. 35, no. 2, pp. 117–135, 2016.

[7] E. J. Sussman, W. G. Tsiaras, and K. A. Soper, "Diagnosis of diabetic eye disease," *Jama*, vol. 247, no. 23, pp. 3231–3234, 1982.

[8] L. Verma, G. Prakash, and H. K. Tewari, "Diabetic retinopathy: time for action. no complacency please!" *BULLETIN-WORLD HEALTH ORGANIZATION*, vol. 80, no. 5, pp. 419–419, 2002.

[9] M. Niemeijer, J. Staal, B. van Ginneken, M. Loog, and M. D. Abramoff, "Comparative study of retinal vessel segmentation methods on a new publicly available database," in *Medical Imaging 2004*. International Society for Optics and Photonics, 2004, pp. 648–656.

[10] J. Staal, M. D. Abràmoff, M. Niemeijer, M. A. Viergever, and B. van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *Medical Imaging, IEEE Transactions on*, vol. 23, no. 4, pp. 501–509, 2004.

[11] Y. Yang and S. Huang, "Image segmentation by fuzzy c-means clustering algorithm with a novel penalty term," *Computing and Informatics*, vol. 26, no. 1, pp. 17–31, 2012.

[12] T. Mapayi, J.-R. Tapamo, and S. Viriri, "Retinal vessel segmentation: A comparative study of fuzzy c-means and sum-entropy information on phase congruency," *International Journal of Advanced Robotic Systems*, vol. 12, no. 1, p. 11, 2015.

[13] T. Mapayi, S. Viriri, and J.-R. Tapamo, "Adaptive thresholding technique for retinal vessel segmentation based on glcm-energy information," *Computational and mathematical methods in medicine*, vol. 2015, 2015.

[14] T. Mapayi and J.-R. Tapamo, "Difference image and fuzzy c-means for detection of retinal vessels," in *2016 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)*. IEEE, 2016, pp. 169–172.

[15] C. Heneghan, J. Flynn, M. OKeefe, and M. Cahill, "Characterization of changes in blood vessel width and tortuosity in retinopathy of prematurity using image analysis," *Medical image analysis*, vol. 6, no. 4, pp. 407–429, 2002.

[16] Q. Li, J. You, and D. Zhang, "Vessel segmentation and width estimation in retinal images using multiscale production of matched filter responses," *Expert Systems with Applications*, vol. 39, no. 9, pp. 7600–7610, 2012.

[17] Y. Yin, M. Adel, and S. Bourennane, "Automatic segmentation and measurement of vasculature in retinal fundus images using probabilistic formulation," *Computational and mathematical methods in medicine*, vol. 2013, 2013.

[18] E. Bribiesca, "A measure of tortuosity based on chain coding," *Pattern Recognition*, vol. 46, no. 3, pp. 716–724, 2013.

[19] E. Bullitt, G. Gerig, S. M. Pizer, W. Lin, and S. R. Aylward, "Measuring tortuosity of the intracerebral vasculature from mra images," *Medical Imaging, IEEE Transactions on*, vol. 22, no. 9, pp. 1163–1171, 2003.

[20] G. Dougherty and J. Varro, "A quantitative index for the measurement of the tortuosity of blood vessels," *Medical engineering & physics*, vol. 22, no. 8, pp. 567–574, 2000.

[21] W. E. Hart, M. Goldbaum, B. Cote, P. Kube, and M. R. Nelson, "Automated measurement of retinal vascular tortuosity." in *Proceedings of the AMIA Annual Fall Symposium*. American Medical Informatics Association, 1997, p. 459.

[22] W. E. Hart, M. Goldbaum, B. Côté, P. Kube, and M. R. Nelson, "Measurement and classification of retinal vascular tortuosity," *International journal of medical informatics*, vol. 53, no. 2, pp. 239–252, 1999.

[23] W. Lotmar, A. Freiburghaus, and D. Bracher, "Measurement of vessel tortuosity on fundus photographs," *Albrecht von Graefes Archiv für klinische und experimentelle Ophthalmologie*, vol. 211, no. 1, pp. 49–57, 1979.

[24] E. M. Martinez-Perez, A. D. Hughes, A. V. Stanton, S. A. Thom, N. Chapman, A. A. Bharath, and K. H. Parker, "Retinal vascular tree morphology: a semi-automatic quantification." *IEEE transactions on bio-medical engineering*, vol. 49, no. 8, pp. 912–917, 2002.

[25] C. M. Wilson, K. D. Cocker, M. J. Moseley, C. Paterson, S. T. Clay, W. E. Schulenburg, M. D. Mills, A. L. Ells, K. H. Parker, G. E. Quinn *et al.*, "Computerized analysis of retinal vessel width and tortuosity in premature infants," *Investigative ophthalmology & visual science*, vol. 49, no. 8, pp. 3577–3585, 2008.

[26] H. Azegrouz, E. Trucco, B. Dhillon, T. MacGillivray, and I. MacCormick, "Thickness dependent tortuosity estimation for retinal blood vessels," in *Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE*. IEEE, 2006, pp. 4675–4678.

[27] E. Grisan, M. Foracchia, and A. Ruggeri, "A novel method for the automatic grading of retinal vessel tortuosity," *Medical Imaging, IEEE Transactions on*, vol. 27, no. 3, pp. 310–319, 2008.

[28] R. Turior, D. Onkaew, B. Uyyanonvara, and P. Chutinantvarodom, "Quantification and classification of retinal vessel tortuosity," *Science Asia*, vol. 39, pp. 265–277, 2013.

[29] R. Turior and B. Uyyanonvara, "Curvature-based tortuosity evaluation for infant retinal images," *Journal of Information Engineering and Applications*, vol. 2, no. 8, 2012.

[30] R. M. Haralock and L. G. Shapiro, *Computer and robot vision*. Addison-Wesley Longman Publishing Co., Inc., 1991.

[31] D. E. Knuth, J. H. Morris, Jr, and V. R. Pratt, "Fast pattern matching in strings," *SIAM journal on computing*, vol. 6, no. 2, pp. 323–350, 1977.

Improving the Energy Efficiency of an Autonomous Power System with Renewable Sources

Inga Zicmane
Faculty of Power and Electrical Engineering
Riga Technical University
Riga, Latvia
Inga.Zicmane@rtu.lv

Kristina Berzina
Faculty of Power and Electrical Engineering
Riga Technical University
Riga, Latvia
Kristina.Berzina@rtu.lv

Tatjana Lomane
Faculty of Power and Electrical Engineering
Riga Technical University
Riga, Latvia
loman@eef.rtu.lv

Konstantins Kasperjuks
Faculty of Power and Electrical Engineering
Riga Technical University
Riga, Latvia
kkasperuks@inbox.lv

***Abstract*—Creating decentralized energy supply systems, there arises the task of providing consumers with electric energy, the parameters of which satisfy the necessary requirements. This problem is especially acute when developing systems based on non-traditional sources using the mechanical energy of renewable natural resources. Taking into account the probabilistic nature of their changes, as well as the significant interaction with each other, the choice of operating modes, types and parameters of energy storage devices for such systems becomes quite a complex and ambiguous task of innovative technologies. The article is devoted to the solution of the problem of developing technical solutions for conjugation of various renewable energy sources ensuring guaranteed electricity supply to autonomous consumers and increasing the energy efficiency of such a system as a whole.**

Keywords—autonomous power supply system, renewable energy sources, energy efficiency, mathematical model, wind station, photovoltaic station, electrochemical generator

I. INTRODUCTION

The urgency of developing and creating promising sources of power supply for autonomous electricity consumers has long been recognized in many industrialized countries. The limited availability of fossil energy sources, as well as the current situation with rising prices for hydrocarbon fuel, along with the exacerbation of the environmental situation are factors stimulating the introduction and use of renewable natural resources in the production of electricity. The use of the latter is inextricably related to the use of systems combining the operation of RES with the electrical network of general purpose. Improving energy efficiency and the need for resource saving are topical problems in almost all countries of the world. To address these problems, new energy policy has been developed in recent years based on the use of hybrid energy systems with renewable energy sources (HRES - Hybrid Renewable Energy System), in which traditional electricity sources are combined with renewable energy sources (such as solar panels, wind generators and others). Such decisions significantly affect the socio-economic factors of the region.

The most important task of designing an autonomous energy system is to choose a rational scheme for its construction, to select and agree on the operating modes of the main power engineering equipment. The task is complicated by the fact that the construction schemes, operating modes and parameters of the power engineering equipment used can differ significantly [1].

Despite the adoption of new regulatory and legislative documents in the field of renewable energy and an expanding range of power plants for RES, the pace of development of local power supply systems using renewable energy in Latvia is slow. One of the barriers preventing the use of RES is the low quality of power supply to autonomous consumers due to the random nature of the supply of renewable energy [2].

Since a guaranteed power source is required for consumers of electricity in decentralized areas, the most promising option was the construction of autonomous systems until recently, wind-diesel and wind-photo-diesel power plants are represented despite the fact that the use of petrol power units and diesel generators is associated with large fuel costs and negative impact on the environment. In addition, when operating this equipment, there is no possibility to optimize their operation under a changing load schedule in order to improve fuel and environmental performance.

One of the ways to solve these problems is the use of combined power plants (CPP), which combine the advantages of renewable energy sources (RES), electric storage devices, as well as units of a new generation operating on hydrocarbon fuel, such as, for example, electrochemical power plant (ECPP) on natural gas or diesel fuel based on molten-carbonate and solid oxide fuel cells, which, in turn, can serve as an alternative to the diesel-driven generator, increasing the energy efficiency of a stand-alone power plant as a whole, and significantly reducing the negative impact on the environment. The fuel for such an electrochemical generator is hydrogen or synthetic gas obtained in terms of thermochemical decomposition of hydrocarbon fuel - reforming.

Thus, for the effective operation of the combined power generation system, a solution of two problems, namely, the coupling of heterogeneous sources and the selection of the parameters of the combined electrical complex based on RES is successful. At the same time, the effectiveness of the use of differently renewable energy sources in combination with an electrochemical generator directly affects the efficiency of local power supply systems, which is expressed in the reliability of electricity supply to consumers, the quality of electric energy supplied to consumers, the capital unit costs per unit of input capacity and the share of consumption of organic fuel [2].

II. BRIEF DESCRIPTION

A. *Prerequisites to a problem solution*

In most practical cases, the problem of finding the optimal solution for building an autonomous power system using RES is multifactorial. Accordingly, its successful solution requires the use of a specialized mathematical tool of multiparametric optimization. At the same time, the use of universal mathematical packages for solving this problem is practically impossible, since it becomes necessary to create specialized databases with parameters of the main power equipment [5, 6], meteorological data arrays, a list of technical limitations, etc. In some cases, with certain assumptions, it is possible to develop its own mathematical tool, which makes it possible to adapt the data of meteorological stations to energy problems. However, the developed mathematical tool requires special skills and, based on the theory of probability, has a guaranteed error [1].

When creating decentralized energy supply systems, there arises the task of providing consumers with electric energy, the parameters of which satisfy the necessary requirements. This problem is especially acute when developing systems based on non-traditional sources using the mechanical energy of renewable natural resources.

The tasks of stabilizing the frequency and voltage of an alternating current of an autonomous power plant are determined by the following factors:

- direct connection of the electric generator to the primary motor; nonstationary nature of the energy flow;
- commensurability of the load power with the power of the generator drive;
- random nature of load change. In power supply systems of stable frequency, the currents and voltages of the electric machine generator change during operation both in amplitude and shape. This circumstance substantially complicates the calculation of the conversion load of the electric machine generator [3].

B. *Adaption to local conditions*

For the territory under consideration, Latvia, active wind activity is characteristic.–The highest wind speed occurs during the winter period. In the winter period, stable south-west winds prevail, in the summer period - winds of the northern component with a predominance of the north-west direction. In the winter months, especially on a change of day and night, the wind is gaining strength. In time this coincides with the daily maximum of electricity consumption.

Due to the instability of climatic conditions, there is an important feature: it is impossible to achieve effective energy saving using only one type of alternative energy sources. To improve energy efficiency, an integrated approach should be applied, based on the proposal to use several, quickly recoverable, highly reliable and efficient sources of electrical energy. This will take advantage of the different types of energy sources to ensure uninterrupted power supply [4].

The parameters of power consumption are given in the form of graphs of electrical loads for each hour with the possibility of adjustment for each month. Density of the distribution of the active load power in the section of the day for each month is shown on Fig. 1.

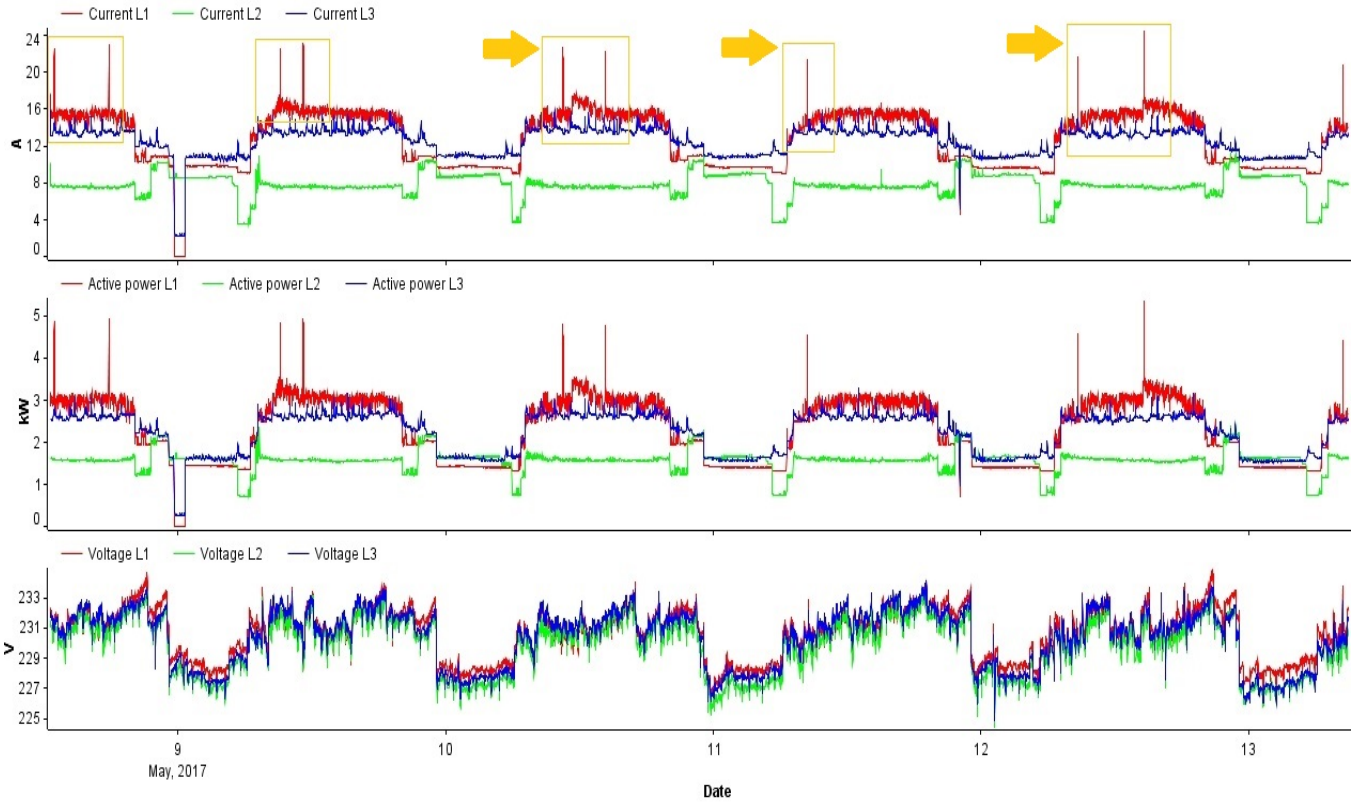


Fig.1. Density of the Distribution of the Active Load Power.

Distribution of the specific solar radiation power by months for February is shown on Fig. 2.

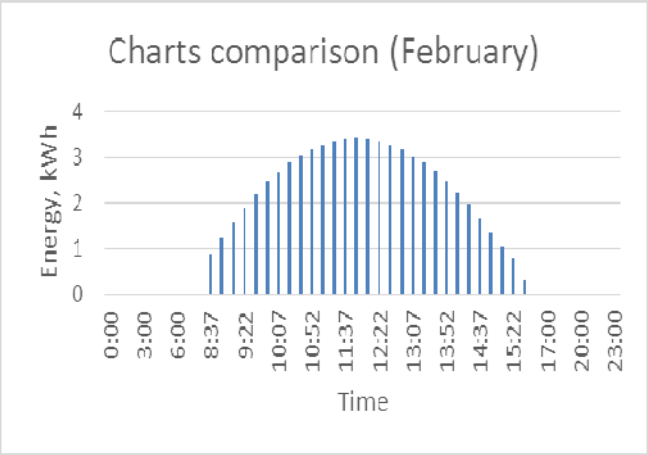


Fig.2. Distribution of the Specific Solar Radiation Power by winter Months.

Probability of distribution of wind speeds in Latvia is given on Fig. 3.

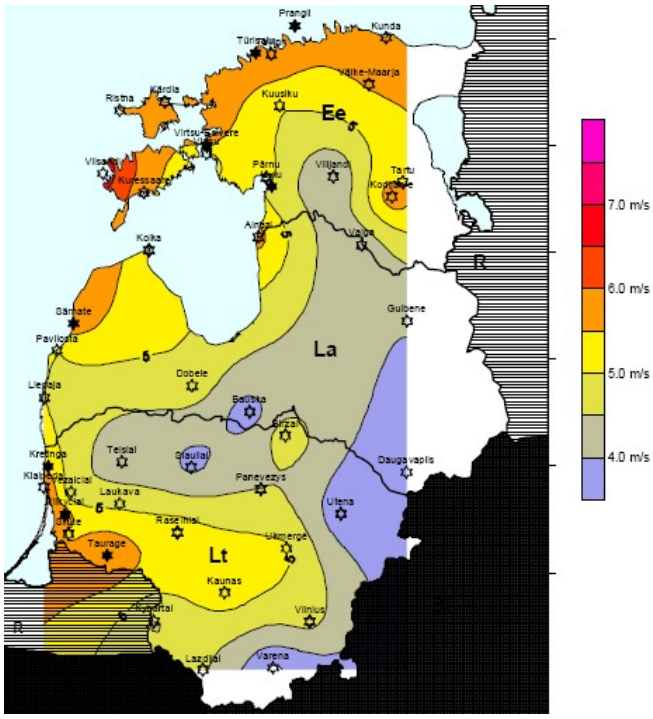


Fig.3. Wind speed in Latvia at 10 meters height [7].

III. Selection of the Circuit of Autonomous Hybrid PS

Hybrid wind-solar systems are used to power a low-power DC load and alternating current in regions where there is sun and wind throughout the year. At the same time, solar and wind energy complement each other (for example, usually winds are blowing in winter, but in summer, there are long sunny days) [6].

Different variants of the coupling between wind power plants (WPP) and photovoltaic power installations (PPI) when working for a general consumer are possible, which can vary significantly both in the composition of the electrical equipment used and in the technical and economic characteristics, of which the most promising option for coupling various types of power plants in one energy system is the use of an intermediate DC insert [5], Fig. 4. This variant of constructing hybrid energy complexes has found wide distribution at small and medium capacities of 1-100 kW, since the application of this approach has great prospects: a hybrid energy complex is built on an aggregate principle, easily scaled and, if necessary, rebuilt. This approach allows unifying the structure and design of electronic power converters: using the modular principle of their construction, it is easier to develop a line of converters for a model range of capacities.

Since the standard household load of 220V AC (lighting, TV, audio system, refrigerator, washing machine, etc.) is used for the electricity supply of residential objects, it is advisable to use a hybrid power system with solar panels and a liquid-fuel generator for more reliable and comfortable power supply, which should include, in addition to solar panels: an inverter, a charge controller and a charger for recharging the accumulator battery (AB) in order to ensure uninterrupted power supply for consumer in calm weather. Also, in order to increase energy efficiency, it is possible to introduce an automatic stop and start-up system of the generator in such hybrid power systems, depending on the degree of charging of the AB [6].

In this circuit of power station construction, consumers are powered by a common autonomous inverter, which ensures high quality of the supplied electric energy. The converter circuits for connecting the PPI and the energy storage device are greatly simplified; a variable-speed WPP can be connected to the system via a controlled rectifier. In addition, the scheme of the station with DC insertion makes it possible to use in its composition the prospective inverter-type DPS [5], which provide significant savings for expensive diesel fuel. At the same time, the power losses associated with the double conversion of electricity by power converters and inverters, given the high efficiency of power electronics, will be insignificant.

IV. Efficienty Criteria of Autonomous Hybrid PS

With this scheme of building a power plant (Fig. 4), the system is easily scaled. Electromagnetic compatibility problems are simply solved and there is no need to reconcile the operation modes of the WPP, PV and electrochemical generator (EG), with each other, which makes it possible to manage these aggregates based on the required optimality criteria determining the efficiency of an autonomous energy complex based on alternative energy technologies: economic and energy. As the only criterion of the economic efficiency of the power generating complex providing the power supply of an autonomous consumer using the conversion of wind energy and solar radiation, it is logical to take the cost of these devices, including their delivery, installation and commissioning (C_{Σ}) [8].

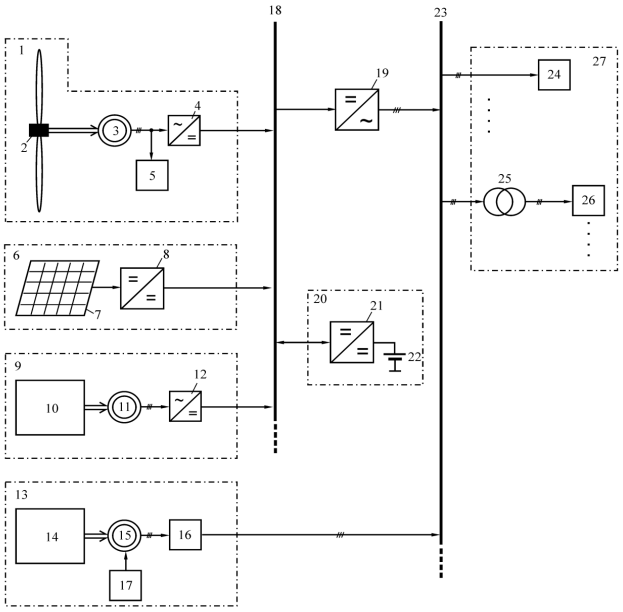


Fig.4. The Scheme of a Hybrid Power Plant with Connection of Generating Units to an Intermediate DC Bus (as well as upon mixed connection) [5].

1 – wind power plant (WPP); 2 – wind turbine; 3, 11, 15 – synchronous electric generator; 4, 12 – controlled rectifier; 5 – block of ballast loads; 6 – photovoltaic power installation; 7 – solar panel (PV); 8 – voltage converter; 9, 13 –generator;10, 14 – electrochemical generator (EG); 16 – soft starter; 17 – field current regulator; 18 DC bus;19 – voltage inverter; 20 – buffer storage of electric power; 21 bi-directional pulse converter; 22 – accumulator battery pack; 23 AC bus 230/400V 50 Hz; 24 electricity consumers; 25 – power step-up transformer; 26 consumers of electricity with voltage of 6 or 12 kV; 27 – decentralized power supply facility

The total cost of the equipment of the autonomous power supply system, taking into account its installation, can be represented in the form:

$$C_\Sigma = C_w \cdot N_w + C_s \cdot N_s + C_a \cdot N_a, \quad (1)$$

where C_w, C_s and C_a – the cost of a wind, photo installation and a storage device, N_w, N_s and N_a, respectively, their quantity.

The values of the parameters N_w, N_s and N_a determine the structure of the power generating complex, that is, the presence or absence of generating devices [8].

The cost of a wind power plant (WPP) depends on its power (the diameter of the wind wheel), the height and design of the mast, as well as includes some of the cost of additional equipment. The cost of a photovoltaic power installation (PPI) is determined by the cost of the panel, taking into account the control systems of its orientation and concentration. In what follows we assume that:

$$C_w = (a_w D_w + b_w) + (c_w h_w + d_w) + \frac{C_{ad} \cdot N_w}{N_w + N_s}, \quad (2)$$

$$C_s = a_s + \frac{C_{ad} \cdot N_s}{N_w + N_s}, \quad (3)$$

where: a_w, b_w, c_w, d_w – empirical coefficients determining the cost of the WPP, depending on its design, power and plant height;

D_w, h_w - diameter of the wind wheel and mast height;

C_{ad} – the cost of additional equipment;

a_s – the cost of one panel, taking into account the control systems of its orientation and concentration [8].

The criteria for energy efficiency include: the maximum total capacity of the generating units of the complex (P_Σ) located in a given territory; the minimum value of the total power dissipated by the ballast resistance ($P_{\Sigma b}$), under the performance condition of the power generating complex.

The choice of the first of the energy criteria as an indicator of efficiency can be used at the stage of performing works related to decision-making on the creation of autonomous enterprises (consumers) in a given territory in a particular geographical point. The application of the second criterion is advisable in situations where the parameters of the consumer's energy costs are determined.

The total power P_Σ in the time interval T:

$$P_\Sigma = \frac{1}{T}\int_0^T (0.5 N_w \rho_w S_w V_w^3 \eta_{ww} \cdot \eta_w + R_s N_s S_s \eta_r \cdot \eta_s) dt \quad (5)$$

where: N_w, N_s – the number of wind turbines and solar panels;

ρ_w – density of air;

S_w – area of the wind wheel;

η_{ww}, η_r, η_s – efficiency of the wind turbine (wind wheel, reducer, generator, etc.), solar panel (concentrator, tracking system, etc.);

η_w, η_s,– efficiency of power transmission lines, communication systems, stabilization, etc., respectively for wind turbines and solar panels;

V_w – wind speed;

R_s – current total solar radiation power (direct, reflected and scattered) in the focusing plane [8].

The total power dissipated by the ballast resistance in the time interval T is determined by the dependence:

$$P_{\Sigma b} = \frac{1}{T}\int_0^T P_b dt. \quad (6)$$

V. Optimization of the APSS

The general idea of the algorithm is to predict the operation of the autonomous power supply system as a response to external influence, while the response has to be determined from the solution of the optimization problem with minimization of the objective function for increments with respect to the optimal solution at the first step of the algorithm implementation and for increments with respect to the dynamic solution at the next steps of the algorithm implementation [9-17]. Figure 5 shows the algorithm of the autonomous power supply system operation.

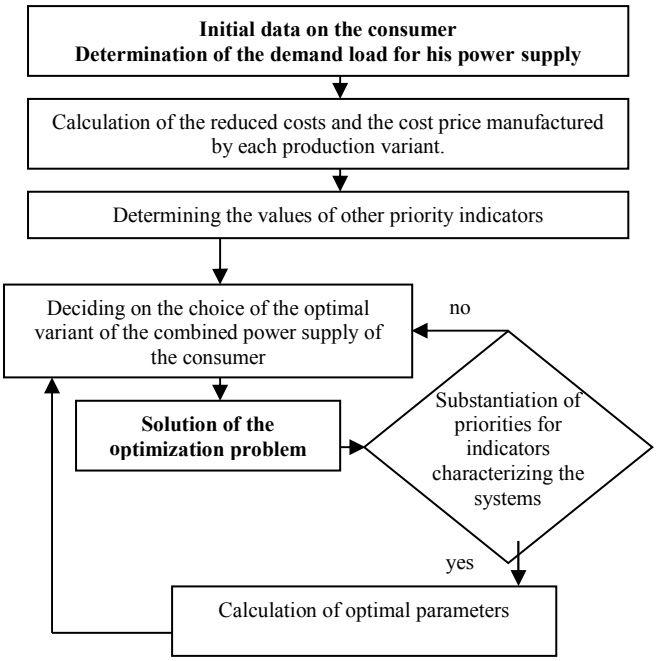


Fig.5. Control-flow chart for optimization of the autonomous power supply system (APSS) operation.

In the future, mathematical and computer modelling of electromagnetic transient processes is provided, taking into account static and dynamic parameters of loads, AC and DC networks of such a system.

VI. Conclusion

The most important task of designing an autonomous power supply system is to choose a rational scheme for its construction, to select and agree on the operating modes of the main power equipment. The task is complicated by the fact that the construction schemes, operating modes and parameters of the energy equipment used can differ significantly.

The idea of sharing the use of WPP, PV and EG is that in the summer months the PV is operating at full capacity, while the WPP at a minimum, in winter - on the contrary. Such an integrated approach in the development of energy systems will improve energy efficiency. This could allow for a more optimal energy consumption and balanced operation, as well as allow using bigger wind turbine generators.

The use of a DC insertion makes it possible to more simply make the summation and distribution of energy flows from generating sources and realize efficient control algorithms for this process. The control system reacts to changing loads and transfers the energy complex to a new optimal state. In the absence of perturbations, the control system returns the state of the energy complex to the initial optimal solution.

References

[1] D.A. Karpov, “Energy efficiency renewable energy small power,” Komsomolsk-on-Amur State Technical University, Russia, pp. 1-2. Д.А. Карпов, “Энергоэффективность возобновляемых источников энергии малой мощности”, Комсомольский-на-Амуре государственный технический университет, Россия, с. 1-2. [Online]. Available: http://www.scienceforum.ru/2014/pdf/993.pdf.

[2] Y.Z. Sarsikeyev, I.I. Sholokhova, S. I. Sheveleva, and A.S. Petrusev, “Technical and economical assessment of stand-alone renewable power system projects,” International research journal, No. 03 (57), Part 4, Yekaterinburg, 2017, pp. 85-90. Е.Ж. Сарсикеев, И.И. Шолохова, С.И. Шевелева, А.С. Петрусёв, “Технико-экономическая оценка проектов электроснабжения в автономных системах на базе возобновляемых источников энергии,” Международный научно-исследовательский журнал, № 03 (57), Часть 4, Екатеринбург 2017, с. 85-90.

[3] A.V. Shalukho, “Increase of efficiency of local power supply systems with renewable energy sources,” The thesis a Cand.Tech.Sci.: 05.09.03 Nizhny Novgorod, 2013, p. 165.

[4] Ngoc Thang, “Management of hybrid energy systems with renewable energy sources,” [Online]. Available: http://www.docme.ru/doc/1434478/upravlenie-gibridnymi-e-nergeticheskimi-sistemami-s-vozobno.

[5] S.G. Obukhov and I.A. Plotnikov, “Comparative analysis of schemes for constructing autonomous power plants using renewable energy facilities, ” Tomsk Polytechnic University, Russia. С.Г. Обухов и И.А. Плотников, “Сравнительный анализ схем строительства автономных электростанций, используя средства возобновляемой энергии”, Томский Политехнический университет, Россия [Online]. Available: http://earchive.tpu.ru/bitstream/11683/27351/1/TPU165314.pdf.

[6] S.A. Burian and N.V. Zykova, “Hybrid power supply system,” Materials of the International Scientific and Practical Conference "Current Issues of Energy Conservation and New Technologies in Science and Education of the Republic of Kazakhstan", No 1, 2014. С.А. Бурьян и Н.В. Зыкова, “Гибридная система электроснабжения”, Материалы Международной Научной и Практической Конференции "Текущие проблемы сохранения энергии и новых технологий в науке и образовании Республики Казахстан", № 1, 2014.

[7] “Wind speed in Latvia” [Online]. Available: http://www.kerveju.lv/ee_veja_karte.php.

[8] V. V. Telegin, "Computer modeling of efficiency of use of systems of alternative power engineering", Natural and technical science, No. 5 (61). 2012, pp. 309-312. В. В. Телегин, "Компьютерное моделирование эффективности использования систем альтернативной энергетики", Естественные и технические науки, № 5 (61). 2012, с. 309–312.

[9] M.P. Kazmierkowski, R. Krishnan, and F. Blaabjerg, “Control in Power Electronics”, Selected Problems. Elsevier Science. USA, 2002, p. 518.

[10] B.V. Lukutin, G.N. Klimova, S.G. Obukhov, E.A. Shutov, and N.M. Parnikov, “Formation of energy-efficient modes of a diesel power plant of an inverter type”, Izvestiya Vuzov. Electromechanics, 2009, No. 6., pp. 80-82. Б.В. Лукутин, Г.Н. Климова, С.Г. Обухов, Е.А. Шутов, и Н.М. Парников, “Формирование энергосберегающих способов дизельной электростанции типа инвертора”, Известия Вузов. Electromechanics, 2009, № 6., стр 80-82.

[11] L.M. Tolbert, W.A. Peterson, M.B. Scudiere, C.P. White, T.J. Theiss, J.B. Andriulli, C.W. Ayers, G. Farquharson, G.W. Ott, and L.E. Seiber, “Electronic Power Conversion System for an Advanced Mobile Generator Set”, IEEE Industry Applications Society Annual Meeting, Chicago, Illinois, September 30 - October 4, 2001, pp. 1763-1768.

[12] K. Berzina, I. Zicmane, and A. Sobolevskis, “Optimal PV electrical energy storage of office building’s communal space lighting”, IEEE International Conference (EEEIC/I&CPS Europe). 2017, pp. 1070-1073.

[13] A.G. Ruiz and M. Molinas, “Electrical Conversion System for Offshore Wind Turbines Based on High Frequency AC Link”, Ecologic vehicles and renewable energies international conference EVER, Monaco, 26-29 March, 2009.

[14] I. Zicmane, K. Berzina, J. Kuckovskis, N. Berzina-Novikova, and I. Zicmane, “The model of places for recreation and relaxation of people taking into account the urboecological aspects of evolution of a Modern City”, 11th International Conference on Electromechanical and Power Systems, SIELMEN 2017. Proceedings, 2017, January, pp. 385-390.

[15] “The combined power systems of autonomous power supply”. “Комбинированные энергетические установки в системе автономного электроснабжения”. [Online]. Available: http://energo.jofo.me/462470.html.

[16] V.M. Gorbov and M.A. Karpov, “The concept of a ship power plant with an electrochemical generator”, Zaporizhzhya National Technical University. No. 2, 2011, pp. 17-24.

[17] K. Berzina, I. Zicmane, T. Lomane, and G. Georgiev, “PV electrical energy storage of residential buildings communal space lighting”, Proceedings of the 9th International Scientific Symposium on Electrical Power Engineering, ELEKTROENERGETIKA 2017, pp. 223-226.

Mathematical Model for Autonomous Power Supply Systems on Renewable Sources for Remote Power Users

Inga Zicmane
Faculty of Power and Electrical Engineering
Riga Technical University
Riga, Latvia
Inga.Zicmane@rtu.lv

Kristina Berzina
Faculty of Power and Electrical Engineering
Riga Technical University
Riga, Latvia
Inga.Zicmane@rtu.lv

Anatolijs Mahnitko
Faculty of Power and Electrical Engineering
Riga Technical University
Riga, Latvia
Inga.Zicmane@rtu.lv

***Abstract*—Trends in the development of the electric power industry in the world are associated not only with the increase in the scale of electricity production in traditional large power plants, but also with the increase in the share of distributed generation [1]. Nowadays, autonomous energy systems have found wide application, in which the electric power generated by installations on renewable energy sources (wind, wave, photoelectric plants, etc.) is given directly to the load (the consumer). However, there are problems in ensuring the quality of electricity caused by instability of the energy flow in the source (wind force, perturbation intensity, light energy flow, etc. are changed), because the distinguishing feature of the windmill electric generating unit is the instability of the output power associated with the variability of the characteristics of the wind as an energy carrier. In these conditions, for reliable power supply of consumers, it is necessary to provide operational control of the operation modes of APS by means of flexible redistribution of active and reactive power, depending on the specific emerging situation associated with changing the load or changing the wind parameters and the corresponding change in the output power of the windmill electric generating unit. The purpose of this paper was to study the issue of increasing the quality of energy of autonomous power supply systems for remote consumers of small and medium power.**

Keywords—autonomous power supply system, renewable energy sources, energy quality, wind station, photovoltaic station

I. INTRODUCTION

Autonomous power supply networks are a wide class of systems used in mobile facilities, technological complexes located in remote regions, as well as in a number of other cases [2]. Recently, autonomous sources of electrical energy are becoming more widespread. The need for them arises where it is technically impossible to use or it is not economically advantageous to use centralized electricity supply, for example, on moving objects [3].

The most significant difference between a system with an autonomous source of electrical energy from a centralized one is that the power of an autonomous source is limited, while a single-power system can be regarded as a source of infinite power. Due to power restriction, the influence of consumers on network performance parameters can be significantly higher than in networks connected to a single power system [2].

To create autonomous power systems of low power that can be used at a distance from power lines and heating lines, as well as in emergency situations, when traditional sources or energy transport systems fail, it is advisable to use renewable energy sources as reserve ones, primarily solar energy radiation and wind flow. However, these energy sources have disadvantages, the main of which are high energy costs, low power, and inconstancy over time during the day and year, unpredictability [4]. As a result of this, the main feature of autonomous systems on renewable sources is a significant dependence of the parameters of electric energy production on random environmental factors (solar radiation intensity), which cannot be influenced by internal means. Such influences, along with internal negative perturbations, such as [5], can be considered as random perturbing factors. They have a different intensity, which in this case will correspond to the depth of the decrease in the intensity of solar radiation as the cloud passes relative to its nominal value in cloudless weather.

Thus, if renewable sources are the primary sources of an autonomous power installation: a mini hydro power plant (mini HPP), a windmill electric generating unit (WEGU), photoelectric converters (FEC) and others, then the power generation of a stand-alone power plant (NPP) depends strongly on the climatic conditions of operation, while the main technical indicators of electric or thermal energy generated by primary RES, such as type of current, frequency and value of output voltage are different. In addition, the energy production mode of RES, as a rule, does not substantially coincide with the energy consumption charts, which, in general, needs electricity, thermal energy, and in some cases, in the cold. Thus, the effective operation and energy balance of an autonomous power plant is determined by the ratio of the change in the energy balance of RES and the graph of the electric load of the system at the consumer. In the process of coordinating the operation of an autonomous power installation on the basis of RES and the consumer, it is necessary to solve the following tasks:

- maximum use of renewable energy resources;.

- distribution of generated and consumed energy, which, as a rule, requires the inclusion of accumulator batteries in the power system;

- regulation by the generated energy parameters [6].

II. BRIEF DESCRIPTION

A. Prerequisites to a problem solution

Ensuring the quality of electricity (QE) for autonomous power supply systems of industrial frequency is a more complicated task than for general purpose networks. At the same time, the deterioration in the quality of electrical

energy has a disastrous effect on many groups of consumers. Existing methods to improve the QE, used for networks of centralized power supply, are not always suitable for autonomous power plants (APP). For example, additional devices that improve QE, lead to a deterioration of the mass-dimensional and value indicators of the autonomous power supply system [3].

So, for example, to smooth out fluctuations of the generated power and to harmonize the latter with the power consumption mode, energy batteries are used: electric batteries, air-accumulating plants, hydroelectric pumped-storage power plants, etc. [7]. However, the energy produced by installations for RES under low wind, solar radiation, waves, etc. is not used, because the quality (harmonicity, voltage, frequency and power) does not satisfy both the operating conditions of the pump motors (the pumps do not work, the electrical coils are out of order) and the consumer, who must continuously receive stabilized electric power from the hydrogenerator.

The PV array and wind turbine work together to satisfy the load demand. When energy sources (solar and wind energy) are abundant, the generated power, after satisfying the load demand, will be supplied to feed the battery until it is full charged. On the contrary, when energy sources are poor, the battery will release energy to assist the PV array and wind turbine to cover the load requirements until the storage is depleted. In order to predict the hybrid system performance, individual components need to be modelled first and then their mix can be evaluated to meet the load demand [8-9].

B. *Prerequisites to a problem solution*

The most promising for use as a power supply system for remote small and medium-sized consumers is currently represented autonomous combined photoelectric and wind power plants, Fig.1.

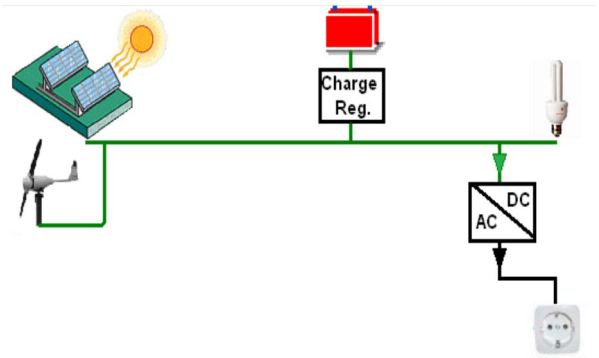


Fig.1. Wind/PV Hybrid system [9].

III. Mathematical Models of NRES

A. *Receiving of solar radiation*

The equation for seasonal effects is an approximation under the assumption of a circular orbit of the earth around the sun. This assumption is allowed because the excentricity is only 0.07. Eq. 1, 2 describes the daily effects and their dependence on the geographical location. [8].

$$\delta_s \approx \Phi_r \cos\left[\frac{C(d-d_r)}{d_y}\right], \quad (1)$$

$$\sin\psi = \sin\phi\sin\delta_s - \cos\phi\cos\delta_s\cos\left(\frac{C\cdot t_{UTC}}{t_d} - \lambda_e\right), \quad (2)$$

where: d_s - solar declination angle, the angle between sun and the plane of the equator [rad];

Φ_r - the tilt of the earth's axis relative the orbital plane of the earth around the sun, $F_r = 0.409$ [rad];

C - constant, $C = 2p$ [rad];

d - day of the year [days];

d_r - the day of the year at summer solstice, 173 (22 June) for non-leap years [days];

d_y - total number of days in one year [days];

f - latitude of the location [rad];

λ_e - longitude of the location [rad];

y - elevation angle [rad];

t_{UTC} - time according to the Coordinated Universal Time [h];

t_d - time of the day [h].

Thus, when creating a mathematical model of sun irradiance, when sky is clear, taking into account the slope ramp:

$$H = \int (E^{ult}\cos\theta + E^{calc} + E_{en})dt, \quad (3)$$

$$\cos\theta = (A-B)\sin\delta + [C\sin\omega + (D+E)\cos\omega]\cos\delta, \quad (4)$$

$$A = \sin\varphi\cos\beta, \quad B = \cos\varphi\sin\beta\cos\gamma,$$

$$C = \sin\beta\sin\gamma,\ D = \cos\varphi\cos\beta,\ E = \sin\varphi\sin\beta\cos\gamma. \quad (5)$$

where:

β - the angle of inclination of the receiving platform to the horizontal plane;

γ - azimuth, the angle between the projection to the horizontal plane of the normal to the surface of the receiver and the meridian (if the surface is directed strictly to the south);

φ - latitude, on which there is a solar receiver;

$\omega = 15(T_s - 12)$ hour angle - the angle of rotation of the Earth from the moment of solar noon.

B. Wind energy

Wind energy is ample, renewable, widely distributed, clean, and works against the greenhouse effect if used to replace the use of fossil-fuel.

Calculation of available Wind generator power is performed with the help of following formulas. The available wind generator power output is a function of the wind velocity. This wind power is calculated for each hour of a typical day in every month: [10]

$$P_w = P_r (V-V_c) / (V_r-V_c) \text{ for } (V_c \le V \le V_r); \quad (6)$$

$$P_w = P_r \text{ for } (V_r \le V \le V_f); \quad (7)$$

$$P_w = 0 \text{ otherwise.} \quad (8)$$

Arrival of the wind flow energy:

$$E = \frac{1}{2} q \bar{v}_w^3. \quad (9)$$

C. Storage model

Currently there are several different storage technologies available. The main types are mechanical (flywheel, pumped hydro etc.), electrical (SMES, capacitors etc.) and electrochemical (batteries etc.) [8]. Which storage technology to use can vary because the applicability of each technology mostly depends on the location? Currently some of the technologies are not sufficiently mature for this application but might be in the future.

Therefore, the behaviour of the storage has been considered ideal in aspects such as efficiency and discharge depth during this study.

Battery storage is sized to meet the load demand during non-availability period of renewable energy source, commonly referred to as days of autonomy [10].

Normally days of autonomy is taken to be 2 or 3 days.

Battery sizing depends on factors such as maximum depth of discharge, temperature correction, rated battery capacity and battery life. Required battery capacity in ampere hour is given by

$$B_{rc} = \frac{E_{c(Ah)} D_s}{(DOD)_{\max} \eta_t},$$

where: $E_{c(Ah)}$ is the load in ampere hour,

D_s is the battery autonomy or storage days;

DOD_{max} is the maximum battery depth of discharge;

η_t is the temperature correction factor [10].

The charge quantity of battery bank at the time t can be calculated by

$$E_B(t) = E_B(t-1)(1-\sigma) + (E_{GA}(t) - E_L(t)/\eta_{inv})\eta_{battery} \quad (11)$$

where: $E_B(t)$ and $E_B(t-1)$ are the charge quantities of battery bank at the time t and $t-1$;

σ is the hourly self-discharge rate;

$E_{GA}(t)$ is the total energy generated by renewable energy source after energy loss in controller;

$E_L(t)$ is load demand at the time t;

η_{inv} and $\eta_{battery}$ are the efficiency of inverter and charge efficiency of battery bank [10].

Charge quantity of battery bank is subject to the following constraints:

$$E_{B_{\min}} \le E_B(t) \le E_{B_{\max}}, \quad (12)$$

where E_{Bmax} and E_{Bmin} are the maximum and minimum charge quantity of battery bank [10].

IV. Selection of non-traditional renewable sources of energy for small and medium power

In order to determine the possibilities of involving an NRES in the energy balance of a particular region, it is necessary to determine with sufficient accuracy their potential on the basis of climate, actinometric, meteorological, and geological and other statistical data available for most of the territories. The complex potential of non-traditional renewable energy sources (NRES) is an additive value that can be estimated for both the studied area as a whole and for any part of it, Fig. 1.

The coefficient of use of wind turbines:

$$\varepsilon_1 = \frac{S_W}{S_W + S_G} \quad (13)$$

where: S_W is full (active P_{Wi} and reactive Q_{Wi}) power generated by the whole set of power turbines:

$$S_{Wi} = \sqrt{(P_{Wi}^2 + Q_{Wi}^2)};$$

n – a number of generating points with a wind turbine,

$$S_W = \sum_i^n S_{Wi};$$

S_G is the same value for the balancer group,

$$S_G = \sqrt{(P_G^2 + Q_G^2)} \text{ [4].}$$

V. Algorithm of operation of the autonomous energy power system (AEPS) based on a photoelectric and wind station

The main requirement for the proposed system presumes constant power balance of the grid, which means that at any sampling interval i, the sum of supplied powers from PV, WT and batteries must be equal to the one being demanded [11-23]. Therefore, the main condition for a properly operating system is the ability to supply sufficient power at all times [11]:

$$P_{PV} + P_{WT} + P_B = P_L + P_{BL} \quad (14)$$

$$P_{RG} = P_{PV} + P_{WT} \quad (15)$$

$$P_{\sum L} = P_L + P_{BL} \quad (16)$$

$$P_{BL} = 0{,}3 \cdot P_L \quad (17)$$

An algorithm describing the operation of the AEPS version, where electricity is provided by wind generators, photoconverters and batteries, is shown on Fig. 2.

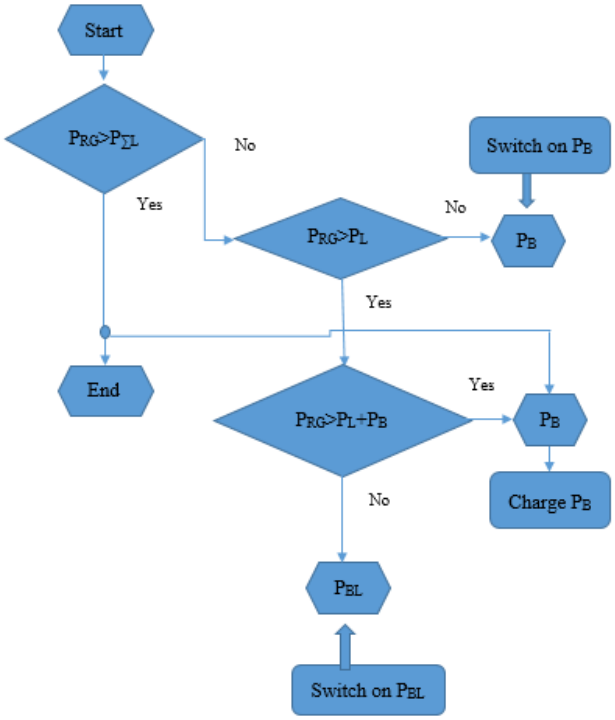


Fig.2. Algorithm describing the operation of the AEPS.

Effective use of the AEPS based on PV and wind EPS with the use of modern automation means involves the development of a number of rules and algorithms for its operation, which should be aimed at improving the energy efficiency of the AEPS. These goals are achieved through careful consideration of all aspects related to the functioning of the system, among which the main ones are:

- the load requested by the consumer at a certain period of time;

- the power produced by the PV and the WG at this moment;

- The power that the system can receive from other sources (battery, DPS) at this moment;

- the cost of electricity from sources included in the AEPS [12].

At the same time, it is necessary to take into account how quickly the power supply can be performed in cases where it is necessary to switch between sources. For example, if the command controller has disabled the DPS, then it may take a certain amount of time to restart it (it can take up to several minutes for the engine to launch and go into operation). Therefore, it is not always possible to get the required amount of energy in the shortest period. Also, diesel generators are not allowed to immediately turn off after a few seconds after starting up.

Exactly on the basis of these and other similar criteria, the basic rules for the functioning of the APS are formed. When modelling the operation of the complex, it is important to take into account all the situations associated with the failure of particular equipment, which theoretically may arise [12].

In the future, mathematical and computer modelling of electromagnetic transient processes is provided, taking into account static and dynamic parameters of loads, AC and DC networks of such a system.

VI. Conclusion

The development of modern AEPS follows the path not only to improve the mass-size, economic, environmental indicators, but also to provide the necessary energy quality. At the same time, with simultaneous growth in the number and quality of development tools, the time allocated for the creation and introduction of new products into production is drastically reduced. The problem of the widespread use of local energy networks (AEPS) based on renewable energy sources is especially relevant for a number of objective reasons, among which there can be called a constant increase in the cost of raw hydrocarbons, an increase in emissions of harmful emissions into the atmosphere, as well as a lack of existing capacities for uninterrupted and high-quality power supply of territories not covered by electric grids and remote from large power plants.

This article suggests algorithm describing the functioning of the AEPS, taking into account different work periods: when the PPS and WPP produce enough power, it is powered by the load and the batteries are recharged (the charge can be full or partial, depending on the battery capacity, the level of solar insolation and wind speed); in other periods of time, the electric power required by the consumer is taken from the accumulator battery.

References

[1] N.I. Voropay, “Distributed generation in power systems,” International scientifically practical conference Small power. 2005.

[2] B. A. Momot, "Reducing the effect of the variable frequency drive of alternating current on the quality of electrical energy in networks with an autonomous source." Б. А. Момот, "Снижение влияния частотно-регулируемого привода переменного тока на качество электрической энергии в сетях с автономным источником," [Online]. Available: http://www.dslib.net/elektrotex-kompleksy/snizhenie-vlijanija-chastotno-reguliruemogo-privoda-peremennogo-toka-na-kachestvo.html.

[3] D.I. Balovnev, "Research and quality management of electrical energy of synchronous generators of autonomous power plants," Д.И. Баловнев, "Исследование и управление качеством электрической энергии синхронных генераторов автономных энергоустановок," [Online]. Available: http://www.dslib.net/elektro-mashyny/issledovanie-i-upravlenie-kachestvom-jelektricheskoj-jenergii-sinhronnyh-generatorov.html.

[4] "Intellectual management of complex technical systems." "Интеллектуальное управление сложными техническими системами," [Online]. Available: http://lc.kubagro.ru/aidos/aidos02/7.8.htm.

[5] A.Shpiganovich and S.V. Dovzhenko, “Analysis of the negative disturbances in autonomous power systems,” News of Tula State University. Technical science. ISSN: 2071-6168, 2016, pp. 33-39. А. Шпиганович и С.В. Довженко, «Анализ негативных возмущений в автономных энергетических системах», Новости Тульского

государственного университета. Технические науки. ISSN: 2071-6168, 2016, стр. 33-39.

[6] “Projection of autonomous road lighting system." “Проектирование автономной системы дорожного освещения," [Online]. Available: http://mirrorref.ru/ref_rnajgernaotrbew.html.

[7] S.V. Zharkov, “Method of operation of a power plant on a renewable energy source in an electric power system,” 1991. С.В. Жарков, “Способ работы энергетической установки на возобновляемом источнике энергии в электроэнергетической системе,” 1991. [Online]. Available: http://www.ntpo.com/izobreteniya-rossiyskoy-federacii/elektroenergetika/alternativnye-istochniki-energii/geotermalnye-volnovye-i-gidroelektrostancii/18263-sposob-raboty-avtonomnoj-energeticheskoj-ustanovki-na-vozobnovlyaemom-istochnike-energii.html

[8] J. Ehnberg, “Autonomous Power Systems based on Renewables on generation reliability and system control,” Göteborg: Chalmers University of Technology , 2007. , p. 64.

[9] Mr.Sandeep V. Karemore, Mr.Subhash, and Y Kamdi, “Optimal sizing of Autonomous Wind-PV hybrid system by genetic algorithm with LPSP technology,” International Refereed Journal of Engineering and Science (IRJES), e- ISSN :2319-183X, p-ISSN 2319-1821, pp.29-35.

[10] J. M. Gonzalez de Durana and O. Barambones, “Object oriented simulation of hybrid renewable energy systems focused on Supervisor control,” IEEE Int. Conf. on Emerging Technologies and Factory Automation Proceedings of the IEEE Int. Conf. on Emerging Technologies and Factory Automation, Palma de Mallorca, 22/09/2009.

[11] V. Ujwala, “Methodology for renewable power generation and optimization of hybrid solar / wind power system using IEEE load pattern,” International Journal of Innovations in Engineering and Technology (IJIET), 2014, pp.295-303.

[12] F. Savrasov, “Variants of the autonomous power supply systems's design with photovoltaic devices and algorithms for their work,” Internet-journal "NAUKOVEDENIE", Issue 6, November - December 2013, pp. 1-13. Ф. Саврасов, “Варианты проектирования автономных систем электроснабжения с фотогальваническими устройствами и алгоритмы их работы,” интернет-журнал «НАУКОВЕДЕНИЕ» Выпуск 6, ноябрь - декабрь 2013 г., с. 1-13.

[13] I. Zicmane, K. Berzina, J. Kuckovskis, N. Berzina-Novikova, and I. Zicmane, “The model of places for recreation and relaxation of people taking into account the urboecological aspects of evolution of a Modern City”, 11th International Conference on Electromechanical and Power Systems, SIELMEN 2017. Proceedings, 2017, January, pp. 385-390.

[14] K. Berzina, I. Zicmane, and A. Sobolevskis, “Optimal PV electrical energy storage of office building’s communal space lighting”, IEEE International Conference (EEEIC/I&CPS Europe). 2017, pp. 1070-1073.

[15] A.G. Ruiz and M. Molinas, “Electrical Conversion System for Offshore Wind Turbines Based on High Frequency AC Link”, Ecologic vehicles and renewable energies international conference EVER, Monaco, 26-29 March, 2009.

[16] “Combined power devices in the system of autonomous power supply”. “Комбинированные энергетические установки в системе автономного электроснабжения”, [Online]. Available: http://energo.jofo.me/462470.html.

[17] V.M. Gorbov and M.A. Karpov, “The concept of a ship power plant with an electrochemical generator,” Publ. Zaporizhzhya National Technical University, No 2 (2011), pp. 17-24. В.М. Горбов и М. А. Карпов, «Концепция судовой электростанции с электрохимическим генератором», Изд. Запорожский национальный технический университет, № 2 (2011), с. 17-24.

[18] K. Berzina, I. Zicmane, T. Lomane, and G. Georgiev, “PV electrical energy storage of residential buildings communal space lighting”, Proceedings of the 9th International Scientific Symposium on Electrical Power Engineering, ELEKTROENERGETIKA 2017, pp. 223-226.

[19] P. Y. Buchatsky, "Development of methods for the analysis and synthesis of energy systems with non-traditional renewable energy sources." П. Ю.Бучацкий, "Разработка методов анализа и синтеза энергетических систем с нетрадиционными возобновляемыми источниками энергии," [Online]. Available: http://www.dslib.net/sys-analiz/razrabotka-metodov-analiza-i-sinteza-jenergeticheskih-sistem-s-netradicionnymi.html

[20] R.A. Dajchman, “Developing of autonomic energy supply”. Privolzhsky Scientific Bulletin No 6 (58) , 2016 11, pp. 11-13.. Р.А. Дайчман, «Разработка системы автономного энергоснабжения». Приволжский научный вестник, № 6 (58), 2016 11, с. 11–13.

[21] B. Wichert, “PV-diesel hybrid energy systems for remote area power generation,” Renewable and Sustainable Energy Reviews, Vol. 1, No 3, 1997, pp. 209-228.

[22] P.Y. Lim and C.V. Nayar, “Control of photovoltaic-variable speed diesel generator battery-less hybrid energy system,” Conference and Exhibition, Manama, 18-22 December 2010, extended abstracts and papers, pp. 223-227.

M. Muselli, G. Notton, and A. Louche, “Design of hybrid-photovoltaic power generator with optimization of energy management,” Solar Energy. Vol. 65, No 3, 1999, pp. 143-157.

Multi-path Load Balancing for SDN Data Plane

M.C. Nkosi, A.A. Lysko, S. Dlamini

CSIR Meraka Institute, Council for
Scientific and Industrial Research
P.O. Box 395
Pretoria 0001, South Africa
{mnkosi2, alysko}@csir.co.za

Abstract— **Networks have become an important feature of our day-to-day to life and therefore, user experience is an imperative goal to be achieved by network operators. Load balancing is a method of improving network performance, availability, minimizing delays and avoiding network congestion. In this paper, we study dynamic load balancing to improve network performance and reduce network response time. The load balancer is applied to OpenFlow SDN network's data plane with Opendaylight as the controller. The flexibility of the load balancer is tested by using it on two different network topologies. Results show that the load balancer can improve the overall performance of the network and reduce delay. The main contribution of this work is a load balancing mechanism for SDN centralized controller environments which can be employed at any point in time in a network, for example, before network failure or after link failure, to avoid data plane congestion and link overloading.**

Keywords—SDN, OpenFlow, Data Plane, Controller, Load balancing

I. INTRODUCTION

Software Defined Networking (SDN) is a key enabling networking paradigm for 5G communications which promises to simplify network management, reduce operational costs, reduce network resource utilization, and stimulate innovation of new services and network evolution [1]. The core function of SDN is to separate the control plane from the data plane by bringing about network control functions based on an abstract representation of the network. The network functions are implemented by removing control decisions, such as routing, from hardware devices and enabling programmable flow tables in the hardware using a standardized protocol, such as OpenFlow (OF). The network abstraction is achieved through the use of a logically centralized SDN controller which defines the behavior of the data plane. When a new flow is initiated in the data plane and there is no routing policy that exists within the flow table, the forwarding device sends the first packet of that flow to the controller. The controller then defines the relevant forwarding route for that flow. One forwarding policy is usually defined for each flow. Two or more forwarding policies maybe defined for a single flow only for backup purposes. Therefore, there are too many packets using a particular flow, that flow may experience overloading and cause delays and the overall network efficiency. Thus, the ever increasing volume of network traffic requires efficient traffic engineering and management methods to ensure availability, scalability and reliability of network.

Load balancing is a method of managing incoming traffic by distributing and sharing the load fairly among available network resources to improve network availability, to reduce latency and bandwidth utilization. In legacy networks, a load balancer is a device designed on a particular hardware [2]. The load balancers are costly and differ based on vendor. In SDN, load balancers are program codes which can easily be implemented on the SDN controller to efficiently manage network load. Most SDN controllers, such as, for example, Onos, Ryu, Pox, floodlight, come with built in static network load balancers with predefined load balancing policies such as, for example, Round Robin, and Random. Mostly, the load balancing in SDN controller is implemented using two approaches, namely, the stateless and the state-full load balancing [3]–[4]. In stateless approach, a controller does not monitor the state of the network and will only load balance traffic at a particular predefined time. This method works well when traffic demands are known in advance. In state-full approach, the controller keeps track of the state of the network and performs load balancing as required. Although state-full approach is preferred, it is expensive because the controller should maintain per-flow states for flow table [5]. This may sometimes lead to controller over loading, which may ultimately cause network unavailability and delays.

In this paper, a dynamic flow load balancer is used to define alternative flows for a flow and then traffic load traversing a flow is shared equally among alternative flows. Load balancing is performed as required to relive the controller from maintaining per-flow states table. Opendaylight is used as the centralized SDN controller. In particular, the objective of this work is to compare the performance of Opendaylight built in load balancer, which is based on round robin and random policies, with a dynamic load balancer. The dynamic load balancer uses Dijkstra algorithm to calculate shortest alternative routes for each flow and balances traffic amongst the defined alternative flows. The flexibility of the load balancer is tested by using different network two different topologies. Mininet emulation tool is used to simulate network topologies, iperf is used to generate traffic load and Wireshark is used monitor the performance of the emulated network.

The rest of the paper is organized as follows. Section two provides related work. Section three presents system setup and methodology. Section four describes results. Section five concludes this paper.

II. RELATED WORK

SDNs, by their nature, accumulate a lot of control traffic due to control plane operations and signaling events that

must be addressed efficiently to ensure effective and resilient networks.vWithout efficient management methods, the control traffic may overwhelm the controller, more particularly in a centralized controller network, and cause delay and loss of information [1],[6]. Work focusing on control plane operations management where controller traffic load balancing is addressed to ensure efficient SDN control plane operation, is discussed in [6]-[7]. Proper management of data plane traffic can relief controller from suffering from overloading. Existing works focus on data plane load balancing to address load balancing after network failures such as link failure, switch failure and less on dynamic load balancing helpful in avoiding network delays and traffic congestion [8]-[9].vWang et al. proposed a path load balancing mechanism which balances traffic after link failure in the data plane [8]. Adami et al, [9], introduced a class-based traffic recovery load balancing method to ensure resilient network. In addition, several recent studies have focused on dynamic load balancing in the data plane from various perspectives [4],[10]-[11].vKhan et. al, studied dynamic load balancing based on traffic volume by monitoring link usage and load balancing traffic among available links to avoid link over loading [4].vIn their method, Floodlight was used as a centralized controller. Gupta et. al studied flow statistics based load balancing using POX controller [10]. In their method, load balancing is performed to avoid server overloading by fairly sharing server connection requests among multiple servers. Mallik et. al, introduced a multi-path congestion control with load balancing to try to protect network from congestions caused by load spikes [11].

All these studies concern load balancing in the data plane using Floodlight or POX as the SDN controller. Different from the existing solutions, we investigate the performance of multipath-based load balancing in a single centralized Opendaylight controller's data plane. The multi-path dynamic load balancing method defines alternative paths for a flow and shares a flow's load among the alternatives paths.

III. SYSTEM SETUP AND METHODOLOGY

Software defined networks (SDNs) consist of a controller and a set of OpenFlow switches. Whenever a new traffic flow request enters a switch, the switch sends a routing request to the controller. After receiving the request message, the controller calculates an optimal path and configures routing tables in all the switches along the optimal path. As shown in Fig. 1, each switch consists of a group table and a set of flow tables with associated action set.

A flow table is made up of flow entries which are flow routes for each source and destination pair. Thus, a single

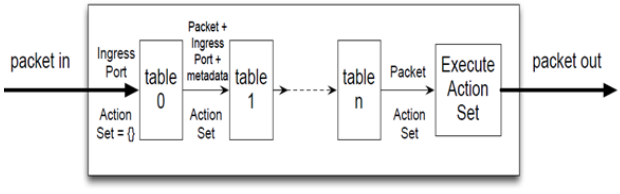


Fig. 1: typical components of an OpenFlow switch

Flow is defined for each source destination pair and all traffic for each pair will use the same route irrespective of how much congested that flow is.

In this study, two scenarios with different network topologies are considered. As depicted in Fig. 2, the first scenario is made up of a network with 7 OpenFlow-switches (OF-switches) and 8 hosts. The second scenario, as shown in Fig. 3, is a bigger network with 40 switches and 81 hosts. The two scenarios were chosen to test the performance, scalability and flexibility of the proposed load balancing method.

Nitrogen version of the open source Opendaylight SDN controller was used as the centralized controller in both the scenarios. Opendaylight is developed by Linux foundation to promote SDN. It is modular, scalable and supports a variety of southbound interfaces such as, other than OpenFlow, the broadband gateway protocol, OVSDB and many others[12]-[13]. Karaf Dlux features [14] were used to monitor the topologies, nodes and controller-switch communications.

Mininet SDN emulation tool was used to emulate the network topologies. Mininet uses a single kernel to run the emulated topologies and employs Open Virtual Switch (OvSwitch) as the default OpenFlow switch[15] . OpenFlow version 1.3 was used because it is still the most supported version in SDN hardware switches.

A source-destination pair was identified in each topology and pings were performed to generate congestion between the source and destination pair. Irrespective of how much the path link flow is congested, new incoming traffic of same source-destination pair is queued on one same data flow path. This at times, may cause delay and ultimately loss of information. To avoid this problem, multi-flow load balancing method was used.

The multi-flow load balancing method calculates alternative short paths which are pushed down into the flow table. A traffic load for a single flow was shared fairly among the alternative flows. The load balancing algorithm takes source and destination pair as an input. The algorithm extracts network topology using JSON and REST APIs and performs link, port, MAC, and IP mappings together with switch and port connections. The algorithm also extracts ports transmission rates statistics to understand the load on each port for each flow. Possible best alternative paths are chosen based on lowest flow cost. Flow cost is calculated as the sum of number of transmitted and received packets at that particular time. PC with Linux Ubuntu 18.04 with 8GB RAM and 2.7GHz processing speed was used to implement this study. IPerf was used to create TCP data streams and to measure the throughput of the data flow before load balancing and after load balancing.

In a nutshell, the following is the step by step implementation methodology for our study for each topology scenario:

- Emulate a network topology using mininet and run mininet *ping all* to ensure that nodes and links are up and running.
- Identify data flow path for a source destination pair.
- Verify that the path is indeed used for transmission for the source-destination pair using Wireshark.

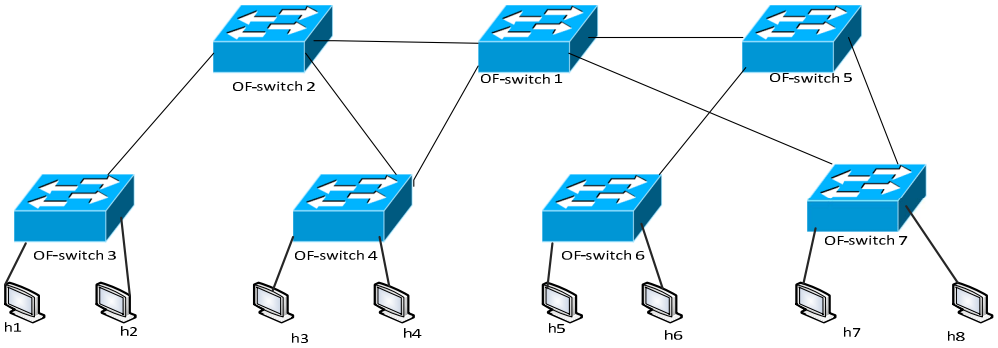


Fig. 2: network topology for scenario 1

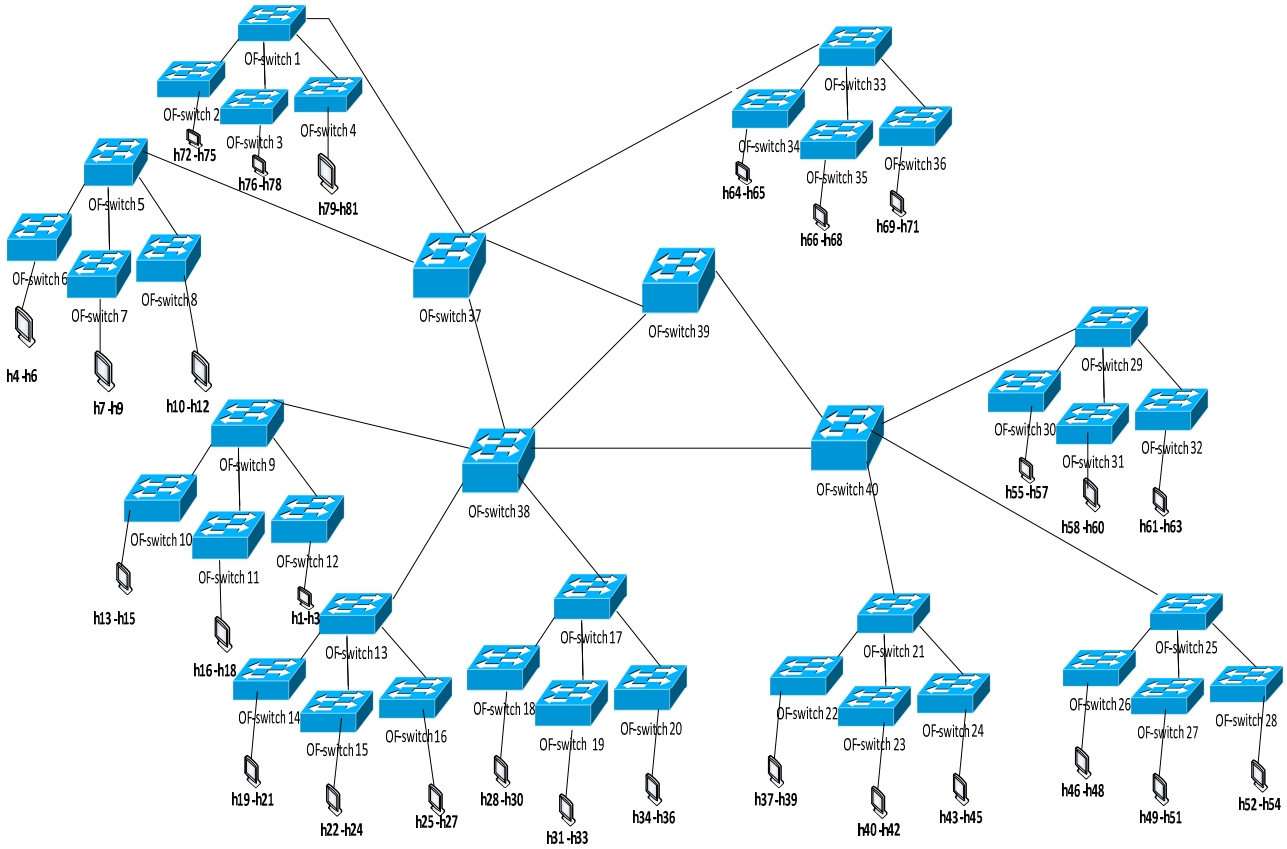


Fig. 3: network topology for scenario 2

- Create a ping on the source-destination pair to generate traffic and flow congestion.
- Perform an iperf and another ping to measure the latency and bandwidth utilization on the over loaded flow. The ping is performed for 10 packets each with packet size 10240 MB. The iperf test was performed using default TCP window frames of size 85.3 KByte with time interval of 15 sec.
- Perform load balancing on the congested source-destination pair (i.e. data path flow)
- Perform an iperf and ping again on the source-destination pair to measure the performance of the load balancing.

Results for both scenarios based on the methodology are highlighted in the next section.

IV. RESULTS AND DISCUSSIONS

A. *Scenario 1*

As shown in Fig. 2, scenario 1 was made of a smaller network with a total of seven switches and eight hosts. A mininet ping all was performed to ensure 100% reachability. H1-h8 was chosen as the source-destination pair. Using Dijkstra's shortest path method, it was calculated that the best path for the pair is [S3-S2-S1-S5-S7]. It was also verified using Wireshark that transmission for the pair use the flow route [S3-S2-S1-S5-S7]. Using mininet xterm, a ping was performed for h1-h8 (i.e., on xterm h1, ping 10.0.0.8 (h1's IP address), and on xterm h8, ping 10.0.0.1(h1's IP address)). After congesting the best path for the source-destination pair, that is the flow [S3-S2-S1-S5-S7], another ping and an iperf with h1 as the client and h8 as the server, were performed to measure bandwidth utilization on the flow before load balancing. The ping results are shown in Fig. 4 and the iperf results are shown in Fig 5.

Load balancing was performed for the source-destination pair. The load balancing algorithm first computes all the flow paths for h1-h8 pair. The following was defined as the paths for the pair h1-h8: [S3-S2-S4-S1-S7], [S3-S2-S1-S5-S7]. The algorithm then computes path costs for all the defined paths by using network statistics. The path cost is calculated as: Cost= Tx +Rx where Tx is number of transmitted packets and Rx is number of received packed. The costs for the defined paths were calculated as [S3-S2-S4-S1-S7:0], [S3-S2-S1-S5-S7:8]. The path with lowest cost was chosen as the shortest path flow. That is to say [S3-S2-S4-S1-S7:1] was chosen and was pushed down to flow routing table as the flow to be used. Load balancing is repeated until path cost for all paths are equal (that is all paths will have same load).

Ping and iperf tests were performed after load balancing to measure bandwidth utilization after the ping. Results are shown in Fig 4 and Fig 5 respectively.

B. Scenario 2

As shown in Fig. 3, scenario 2 was made of bigger network with a total of 40 switches and 81 hosts. Same process as described in section iv.A, was followed for scenario 2. The *h7-h54 source*-destination pair was used with the path flow *[S5-S37-S39-S40-S25]* as the defined default flow by Opendaylight controller. Ping and iperf results are shown in Fig. 6 and Fig. 7 respectively.

For load balancing, the following were defined as the possible path flow with associated costs:

*[S5-S37-S39-S40-S25: **4**]; [S5-S37-S38-S40-S25:**0**];*

*[S5-S37-S38-S39-S40-S25:**0**];[S5-S37-S39-S38-S40-S25:**0**].*

Path *[S5-S37-S38-S40-S25]* was chosen as the new path and load balancing was repeated until load was fairly distributed among the alternative path flows. Ping and iperf results are also shown in Fig. 6 and Fig. 7 respectively.

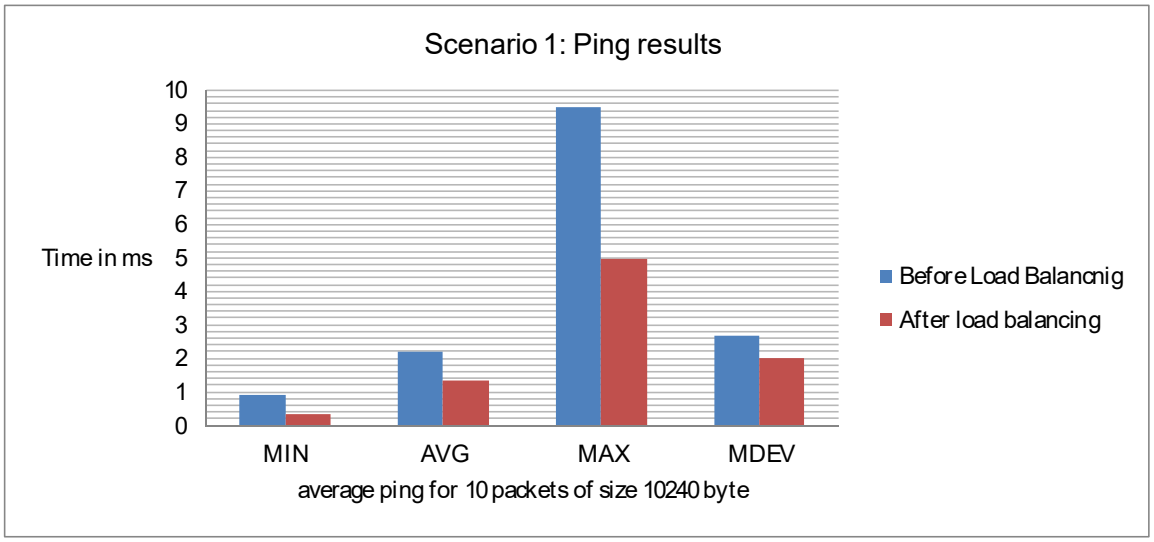


Fig. 4: Scenario 1 ping results for source-destination pair: h1-h8

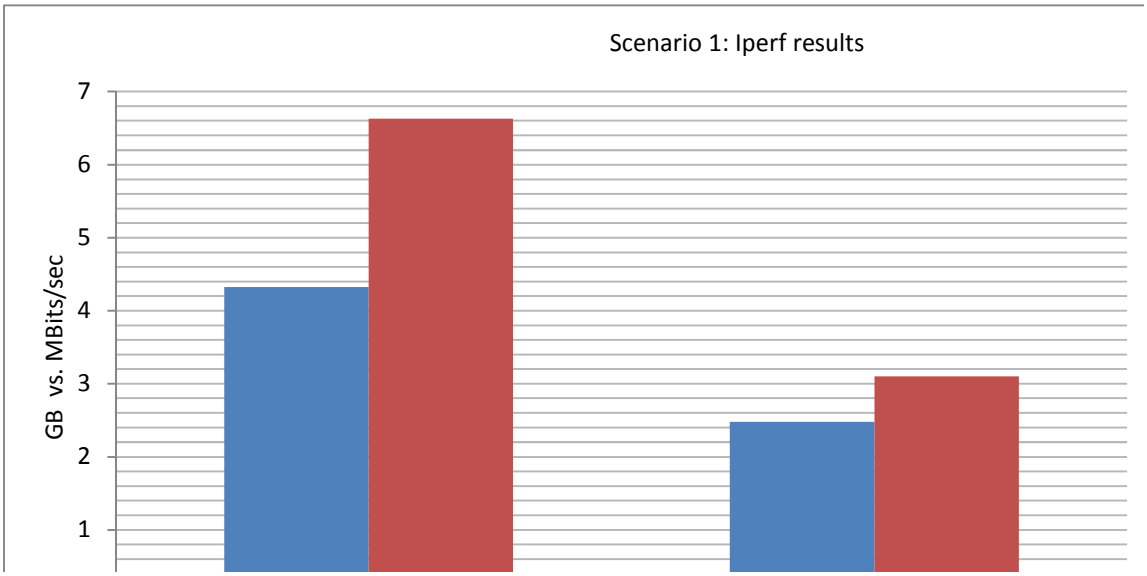


Fig. 5: Scenario 1 Iperf results for source destination pair: h1-h8

For load balancing, the following were defined as the possible path flow with associated costs:

*[S5-S37-S39-S40-S25: **4**]; [S5-S37-S38-S40-S25:**0**];*

*[S5-S37-S38-S39-S40-S25:**0**]; [S5-S37-S39-S38-S40-S25:**0**]*.Path *[S5-S37-S38-S40-S25]* was chosen as the new path and load balancing was repeated until load was fairly distributed among the alternative path flows. Ping and iperf results are also shown in Fig. 6 and Fig. 7 respectively.

C. *Discussions*

As it can be observed from the results of scenario 1, the maximum average ping after load balancing is 4.9m/s as compared to that of before load balancing which is 10m/s. The iperf results show an increase of about 2.2GB in data transfer after load balancing. However, the increase is not as significant as expected. This is assumed to be due to the fact that part of the alternative path routing path uses some link from the congested path. Notice that, the first links in both paths are the same: [S3-S2-S1-S5-S7] [S3-S2-S1-S5-S7]. From scenario 2, it can be observed that the increase is significant in data transfer, from about 4GB to 10 GB.

This is because, unlike, in scenario 1, there are multiple different routing paths that the load balancer can use to transfer data. Therefore, the load balancer works much quicker when there are more options in alternative paths.

From scenario 1 and 2, it can be concluded that the load balancer is flexible for both smaller networks and larger networks. It also can improve network performance and avoid overall network delay.

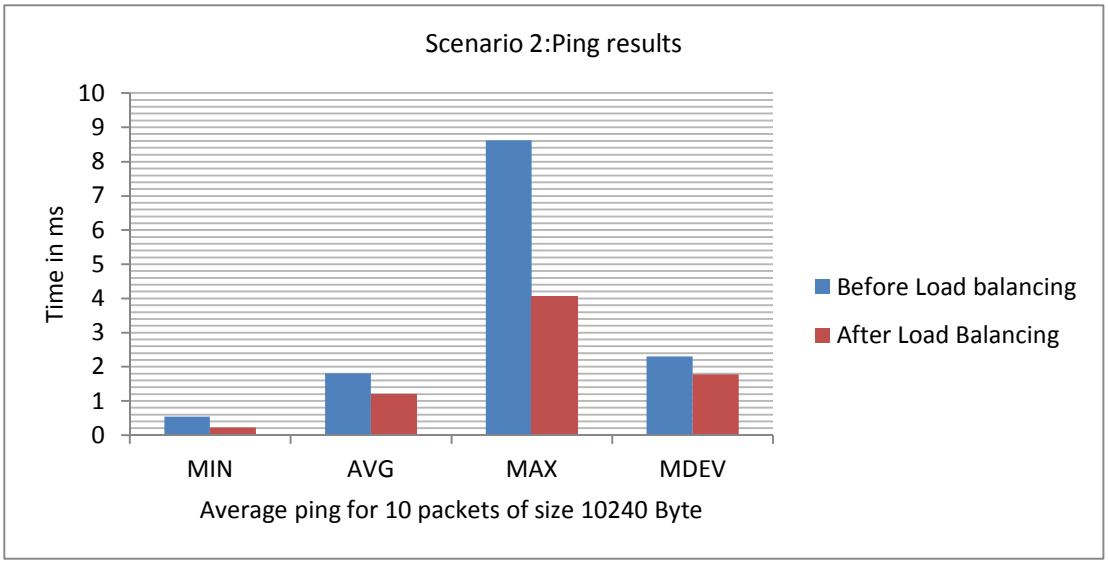


Fig. 6: Scenario 2 ping results for source-destination pair: h7-h54

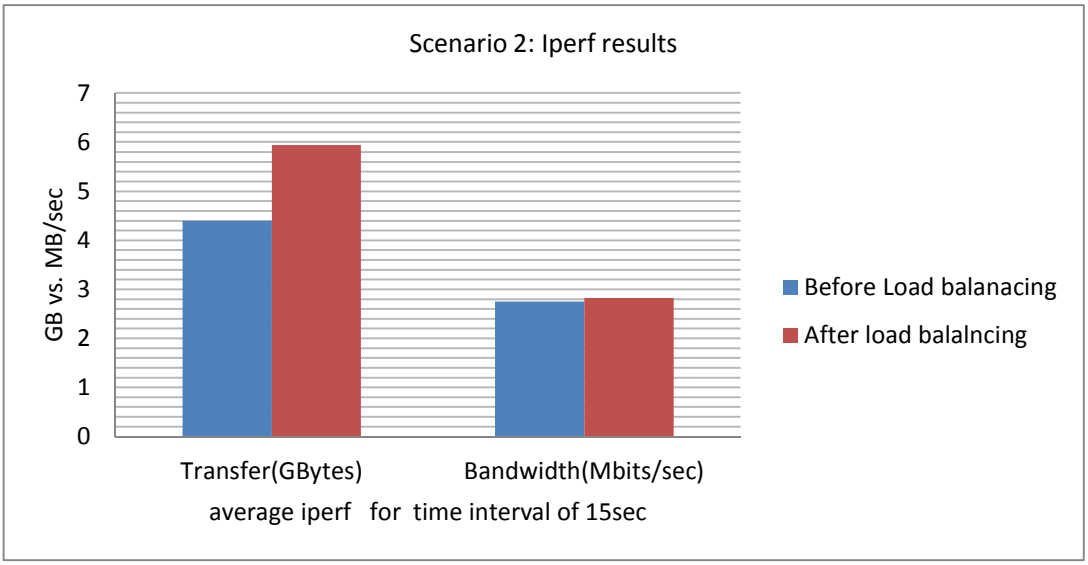


Fig. 7: Scenario 2 Iperf results for source destination pair: h7-h54

V. CONCLUSION

We have performed load balancing for both larger and smaller networks using a dynamic multi-path load balancer in an OpenFlow based SDN data plane. The load balancer determines alternative paths first and then reroute traffic equally amongst the defined paths. The network topologies were emulated using mininet emulation tool. The network uses SDN Opendaylight controller which uses built in load balancer based on round robin and random policies. The performance of the built in load balancer was compared to the multi-path load balancer method. The load balancer has improved the overall network performance in transfer rate and response time. However, it was found that for better network improvement, the data plane should have multiple alternative links so that multiple path paths can be defined for a routing path. The overall contribution of this work is a multi-path load balancing method, which, unlike other load balancing method, can be applied to a network at any state(before date plane failure, or after data plane failure), to ensure network efficiency.

REFERENCES

[1] G. Bianchi, A. Capone, M. Bonola, G. Bianchi, and M. Bonola, "Public Review for OpenState : Programming Platform-independent Stateful OpenFlow Applications Inside the Switch a c m s i g c o m m OpenState : Programming Platform-independent Stateful OpenFlow Applications Inside the Switch," vol. 44, no. 2, pp. 44–51, 2014.

[2] W. K. Soo, T.-C. Ling, A. H. Maw, and S. T. Win, "Survey on load-balancing methods in 802.11 infrastructure mode wireless networks for improving quality of service," *ACM Comput. Surv.*, vol. 51, no. 2, 2018

[3] S. Bhandarkar and K. A. Khan, "Load Balancing in Software-defined Network (SDN) Based on Traffic Volume," vol. 2, no. 7, pp. 72–76, 2015

[4] "US9621642B2 - Methods of foMethods of forwarding data packets using transient tables and related load balancers."

[5] S. H. Yeganeh, A. Tootoonchian, and Y. Ganjali, "On scalability of software-defined networking," *IEEE Commun. Mag.*, vol. 51, no. 2, pp. 136–141, 2013.

[6] T. Kim, S. G. Choi, J. Myung, and C. G. Lim, "Load balancing on distributed datastore in opendaylight SDN controller cluster," *2017 IEEE Conf. Netw. Softwarization Softwarization Sustain. a Hyper-Connected World en Route to 5G, NetSoft 2017*, 2017

[7] J. Li, X. Chang, Y. Ren, Z. Zhang, and G. Wang, "An effective path load balancing mechanism based on SDN," *Proc. - 2014 IEEE 13th Int. Conf. Trust. Secur. Priv. Comput. Commun. Trust. 2014*, pp. 527–533, 2015

[8] D. Adami, S. Giordano, M. Pagano, and N. Santinelli, "Class-based traffic recovery with load balancing in software-defined networks," *2014 IEEE Globecom Work. GC Wkshps 2014*, pp. 161–165, 2014.

[9] K. Kaur, S. Kaur, and V. Gupta, "Flow statistics based load balancing in OpenFlow," *2016 Int. Conf. Adv. Comput. Commun. Informatics*, pp. 378–381, 2016

[10] A. Mallik and S. Hegde, "A novel proposal to effectively combine multipath data forwarding for data center networks with congestion control and load balancing using Software-Defined Networking Approach," *2014 Int. Conf. Recent Trends Inf. Technol. ICRTIT 2014*, 2014

[11] Z. K. Khattak, M. Awais, and A. Iqbal, "Performance evaluation of OpenDaylight SDN controller," *Proc. Int. Conf. Parallel Distrib. Syst. - ICPADS*, vol. 2015–April, pp. 671–676, 2014.

[12] P. Berde *et al.*, "Facilitation of the OpenDaylight Architecture," *Proc. third Work. Hot Top. Softw. Defin. Netw. - HotSDN '14*, pp. 1–6, 2014.

[13] S. Badotra, "Open Daylight as a Controller for Software Defined Networking," vol. 8, no. 5, pp. 1105–1111, 2017.

[14] N. Handigol, B. Heller, V. Jeyakumar, B. Lantz, and N. McKeown, "Reproducible network experiments using container-based emulation," *Proc. 8th Int. Conf. Emerg. Netw. Exp. Technol. - Conex. '12*, p. 253, 2012.

[15] S. Badotra, "Open Daylight as a Controller for Software Defined Networking," vol. 8, no. 5, pp. 1105–1111, 2017.

The Design of Byzantine Attack Mitigation Scheme in Cognitive Radio Ad-hoc Networks

Sekgoari Mapunya
Dept. of Computer Science
University of Limpopo
Polokwane, Sounth Africa
sekgoari.mapunya@ul.ac.za

Mthulisi Velempini
Dept. of Computer Science
University of Limpopo
Polokwane, South Africa
mthulisi.velempini@ul.ac.za

***Abstract*—The ever-increasing number of wireless network systems brought a problem of spectrum congestion leading to slow data communications. All of the radio spectrums are allocated to different users, services and applications. Hence studies have shown that some of those spectrum bands are underutilized while others are congested. Cognitive radio concept has evolved to solve the problem of spectrum congestion by allowing cognitive users to opportunistically utilize the underutilized spectrum while minimizing interference with other users. Byzantine attack is one of the security issues which threaten the successful deployment of this technology. Byzantine attack is compromised cognitive radios which relay falsified data about the availability of the spectrum to other legitimate cognitive radios in the network leading interference. In this paper we are proposing a security measure to thwart the effect caused by these attacks and compared it to Attack-Proof Cooperative Spectrum Sensing.**

Keywords—Radio spectrum, byzantine attack, cognitive radio

I. INTRODUCTION

Slow data communication is not desired unfortunately, with the ever-increasing demand of broadband wireless technologies and services, spectrum bands are getting congested dramatically affecting data transmission speed. spectrum measurements in [1], proved that the is an abundant portion of spectrum that is underutilized. Cognitive radio network (CRN) was first proposed by Joseph Mitola, the proposed technology solve the problem of spectrum congestion by allowing cognitive radios/secondary users to opportunistically utilize the unutilized radio spectrum [2]. In CRN we have two types of users: primary users(PU) and secondary users (SU) which are licensed and unlicensed users of radio spectrums respectively. This technology can be deployed in an infrastructure or infrastructure less network environment. In this paper we are focusing on an infrastructure-less network environment where we are not having any assistance from a base station. Spectrum sensing is the heart of this technology, since SUs, must have full knowledge about the radio environment before accessing the available spectrum to avoid PU interference. However, due to lack of global knowledge, one cognitive node cannot accurately conclude about the availability of the spectrum. This is further worsened by the hidden node problem, false alarm, and inaccurate signal detection. Hence cooperative spectrum sensing (CSS), a form of spectrum sensing where cognitive nodes take decisions about the availability of spectrum cooperatively to avoid wrong decisions is preferred. However, CSS is vulnerable to attacks, which can dramatically affect the performance of the network. Byzantine attack is one of the security issues, which degrades the performance of CRN when CSS is considered [3], [4]. Byzantine attack sometimes known as spectrum sensing data falsification (SSDF) attack mislead the entire network by relaying false information about the radio spectrum occupancy to all the nodes in the network, hence degrading the performance of the network [5].The central premise for such behaviors is either, the attacker is greedy or malicious. greedy attacker sends reports indicating that the spectrum is occupied while not. This causes SU to stop using the channel and the attacker use the channel. This behavior leads to increase in probability false alarm. When the attacker is malicious, it declares that the channels are not occupied while they are occupied to make the SU access the channels and cause interference to the primary users. This action leads to increase in the misdetection probability. Success probability is the probability of correctly detecting malicious users.

In this paper we propose the byzantine attack mitigation scheme known as Extreme Studentized Consensus Cooperative Spectrum Sensing(ESCCSS) scheme that extremely combat the effect caused by greedy attacker known as always yes and malicious attackers known as always No. the effect is combated by isolating falsified data from the final decision about the spectrum occupancy we compared the results of ESCCSS scheme with Attack-Proof Cooperative Spectrum Sensing(APSCC) [6] to check which scheme performs the best looking at success probability.

Contributions of this paper are as follows:

1. We contribute to detection and combatting the effect caused by byzantine attack.
2. Contribute to the body of knowledge in the field of cognitive radio ad hoc networks security.

The rest of this paper is organized as follows: The related work is presented in Section II. In Section III, we describe the proposed scheme. The results of effectiveness of our proposed scheme is presented in Section IV. Finally, we conclude and proposed future work in Section V and VI respectively.

II. RELATED WORK

Literature has shown that less work has been done in decentralized CRN. We analysed a few of such works. APSCC is proposed in [6], to combat the SSDF attacks, some countermeasures are used in the iterative process of information sharing. This enables most legitimate SUs to prohibit potential malicious users (MU) to share data with themselves. simulation results proved that the proposed scheme alleviates different SSDF attacks than several existing schemes.

Lavanis and Jaliahal in [7] evaluated the performance of various algorithms, which are used on a fusion centre, and none of these were considered to be effective in addressing the effects of MU. Fusion center is just a dedicated central point where global decision about the radio environment is reached.

Standard and p-nom energy detector algorithm were computed under the assumption of a Gamma and Gaussian distribution of the test statistic (TS). Performance comparison of these algorithms indicated that the Gaussian assumption for distribution of test statistic were suitable in the setup where SSDF attack is injected as compared to the Gamma assumption. It is assumed that the percentage of MU is significantly more less in the network, how will the algorithm perform when having more MU in the network.

In [8], a scheme designed to counter the SSDF attack based on hard decision was proposed. The proposed scheme identifies the normal users by using Gaussian distribution of Binomial distributions and prevent the malicious user from participating in cooperative spectrum sensing. The authors assumed that the network was populated with 50 nodes. In our work, we simulate our proposed scheme on a network having less than 50 and more nodes.

In [9], authors proposed a scheme, which is based on the prior information of the cognitive radios. It detects and isolate MUs. The proposed scheme uses utility optimization to combat SSDF attack, which can propel the attack even when the malicious nodes cannot be identified. It is also proven that by using the weighted sensing fusion mechanism, CRN can reduce the effects of MUs.

Srinu and Mishra in [10], proposed an algorithm based on extension of generalized extreme studentized deviate (EGESD) test to isolate the malicious nodes in the network. The results show that the EGESD addresses the limitation of the generalized extreme studentized deviate test and that it is superior in addressing the effects of MUs. The scheme is optimized for centralized CSS.

In [11], the use of cognitive radio networks for smart home energy management is proposed. The proposed solution was evaluated to assess its reliability and efficiency. The results show that 91.42% of reliability can be achieved. The authors also assumed the always-on attack and other types of attack behaviour were not considered.

Consensus-based cooperative spectrum sensing scheme to is proposed in [12], counter attack SSDF in CR-MANETs. The scheme is based on recent advances in consensus algorithms [13], that which was magnified by a self-organizing behaviour of animal groups such as fish. Unlike the existing schemes, common receiver is not needed to counter SSDF attacks by fussing the data. trustworthiness of spectrum sensing terminals can be tested by the use of consensus state of secondary users, the proposed scheme can differentiate the falsified data and the legitimate data, which makes it more resilient against SSDF attacks.

III. proposed scheme

The extreme studentized deviate test was implemented to address the effects of byzantine attack in CRN. Consensus algorithm was also used to enable SUs to share and arrive at a global decision about the spectrum occupancy. Hence ESCCSS scheme is a consensus cooperative spectrum sensing scheme. SUs cooperatively arrives at a global decision without the assistance of a base station. Each SU exchange actual spectrum observation with its neighbours for K time steps. Upon the termination of the algorithm, each SU individually makes a decision of H1 or H0 denoting the absence or the presence of PU respectively, based on the received observations from neighboring nodes.

In our consensus based system, SUs cooperates in sensing the spectrum band using energy detection as in fig.1. There are wide spectrum sensing technics used for spectrum sensing those include matched filter, energy detection etc. [14]. we chose, energy detection because of its widely usage in research, low complexity and it does not require any prior knowledge about the primary signal.

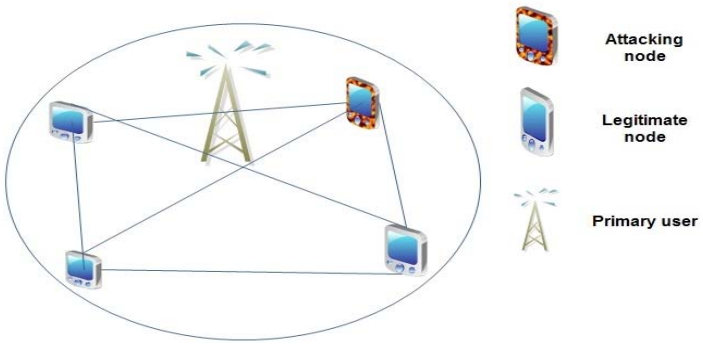


Fig. 1. Distributed cooperative spectrum sensing in the presence of byzantine attack

Fig 1. show an ad hoc network environment populate with four cognitive radios sharing their spectrum observations with each other in the presence of malicious radio. ESCCSS scheme allows cognitive radios to share actual energy detected values, unlike other schemes where secondary users compute and share binary values [15]. Each cognitive radio compute $X_i(n)$ as the average of all the observations at each time step j after the sharing of the data, isolation of malicious data and consensus algorithmic has run. The final decision about the spectrum occupancy will be done using the following equation:

$$X_i(n) = \frac{1}{N}\sum_{j=1}^{N} y_j(n) \quad (1)$$

Where N is the maximum time steps at which each SU observe and record energy values, n is the number of node and $y_j(n)$ is the observed energy value at any time step j and i is the node index.

The computed average is going to be compared to a threshold value to make a final decision.

$$Decision = \begin{cases} 1; & X_t(n) \geq \beta \\ 0; & otherwise \end{cases} \quad (2)$$

If the average is greater than the threshold β then the spectrum is said to be in use denoted by 1, otherwise the spectrum is said to be idling and secondary users can make use of the spectrum.

Based on the key component our opposed scheme can be described in Algorithm 1.

Algorithm 1

1: Sort the received energy values $y_1, \dots y_N$ of N SUs at time k in ascending order.

Let this sorted value be denoted by $x_1, \dots x_N$

2: Estimate the number of outliers/malicious users U

3: Compute the $\bar{x}$ mean and s standard deviation of the received energy values $y_1, \dots, y_N$

4: Compute $R_j = max_i \left\{ \frac{|x_i - \bar{x}|}{s} \right\}$ j = 1, 2,…U

5: After computing R_j Put x_i aside that has maximized $|x_i - \bar{x}|$

6: Repeat step 1 to step 5 with estimated outliers been removed, up until j= U

7: Declare isolated x_i 's as suspicious data and they are excluded from participating when consensus algorithm is run.

Figure .2 present the flow chart for spectrum occupation desertion making which has to be taken by cognitive users. The proposed scheme and consensus algorithm both are activated after the information sharing has occurred.

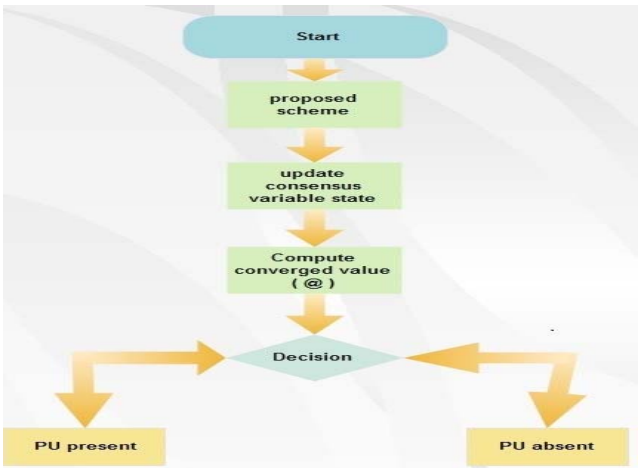


Fig. 2 flow diagram of the process for spectrum occupation decision

When the network is running and cognitive users are taking decisions whether the spectrum is available for use or not, figure .2 shows how each SU takes desertion about the availability of the spectrum in the presence of the proposed scheme. Each SU start by sensing the radio spectrum and share its observations with other nodes in the network. ESCCSS is activated immediately after the sharing of data and the suspected falsified data is alleviated from the decision making. Consensus algorithm is activated and the shared data is combined so that each node in the network arrives at a consensus stage. Thereafter the average of the consensus value is compare to a threshold value to determine whether the spectrum is available or not.

IV. RESULTS

MATLAB 2015a installed in WINDOWS 8.1 operating system was utilized to simulate the effectiveness of the proposed against APCSS Scheme. The simulation was run for 100 seconds on an area of 1000 square meters. We considered 5 scenarios, the first scenario we simulated the scheme in an environment having 10 nodes. In the second to fifth scenario, we had 15, 25, 50, 100 nodes respectively. Each scenario, a case with 10% and 15% of MUs were considered. The effectiveness of ESCCSS scheme where evaluated based on false alarm probability in all the cases. We compared our proposed scheme with APCSS as they are both designed to work on the same network environment. The results are depicted in Figure 3 to 4.

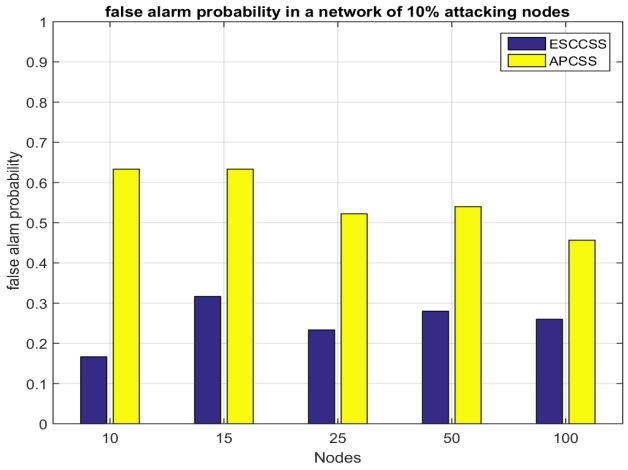


Fig. 3. The False Alarm probability with 10% of malicious users in the network.

Figure .3 presents the results of false alarm probability in all the scenarios with 10% of the nodes being MUs. The results show that as the number of nodes increases the probability of false alarm decrease for APCSS but it increases when the network size reaches 100. It is also observed that as the number of nodes increases there is no constant change of FAP in the performance of the proposed scheme. The scheme significantly reduces FAP compared to APSCC in all the scenarios. Fig 4 shows the results of the scenario with 15% MUs.

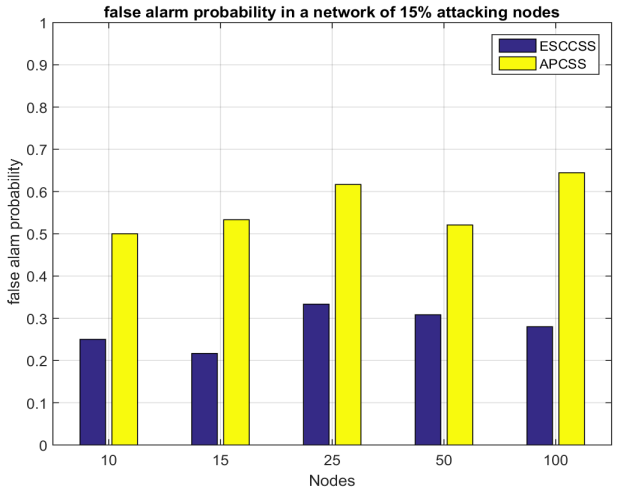


Fig. 4. The False Alarm probability with 15% of malicious users in the network.

As we increased the number of attacking nodes (15%) in fig.4, the proposed scheme managed to address significantly the effects of the byzantine attack by reducing the FAP extensively compared to APCSS. In all the scenarios as the percentage of the MUs is constant the proposed scheme report at most the same FAP while APCSS report a significant difference.

V. CONCLUSION

This paper investigated byzantine attack in cognitive radio ad-hoc networks and proposed a scheme which addresses the effects of these attacks by isolating falsified reported data to all the cognitive users in the network. the use of extreme studentized deviate test and consensus algorithm was proposed to counter the effects of the attacks and allow the smooth running of the network. The proposed scheme was simulation and compared to APCSS, the simulation results showed that ESCCSS which is the proposed scheme significantly reduce the false alarm probability which is desired. Therefore, the proposed scheme outperforms APCSS in terms of FAP. In the future the proposed scheme and APCSS will be evaluated and compared based on other metrics to conclude confidently on its performance.

REFERENCES

[1] M. T. Masonta, M. Mzyece and N. Ntlatlapa, "Spectrum Decision in Cognitive Radio Networks:A suvey," IEEE Communications Surveys & Tutorials, vol. 15, no. 3, pp. 1088-1107, 2012.

[2] J. Mitola, "Cognitive Radio - An Integrated Agent Architecture for for software defined radio," 2000.

[3] L. Zhang, G. Ding, F. Song and Q. Su, "Defending against Byzantine attack in cooperative spectrum sensing relying on a reliable reference," in 2016 IEEE/CIC International Conference on Communications in China (ICCC), Chengdu, 2016.

[4] A. Vosoughi, J. R. Cavallaro and A. Marshall, "A Cooperative Spectrum Sensing Scheme For Cognitive Radio Ad Hoc Networks based on Gossip And Trust," in IEEE Global Conference on Signal and Information Processing (GlobalSIP), Atlanta, GA, 2014

[5] S. Mapunya and M. Velempini, "Investigating spectrum sensing security threats in cognitive radio networks," in *Ad Hoc Networks*, Niagara , Springer, 2018, pp. 60-68.

[6] Q. Liu, J. Gao, Y. Guo and S. Liu, "Attack-Proof Cooperative Spectrum Sensing Based on Consensus Algorithm in Cognitive Radio Networks," KSII Transactions on Internet and Information Systems (TIIS), vol. 4, no. 6, pp. 1042-1062, 2010.

[7] N. Lavanis and D. Jalihal, "Performance of p-Norm Detector in Cognitive Radio Networks with Cooperative Spectrum Sensing in Presence of Malicious Users," Wireless Communications and Mobile Computing, vol. 2017, no. 2, pp. 1-8, 2017.

[8] J. Lu, P. Wei and Z. Chen, "A Scheme to Counter SSDF Attacks based on Hard Decision in Cognitive Radio Networks," WSEAS TRANSACTIONS on COMMUNICATIONS, vol. 13, pp. 242-248, 2014.

[9] S. Sodagari, A. Attar, V. Leung and S. G. Bilén, "Combating channel eviction triggering denial-of-service attacks in cognitive radio networks†," transactions on emerging telecommunications technologies, vol. 23, no. 5, pp. 454-465, 2012.

[10] S. Srinu and A. K. Mishra, " Efficient elimination of erroneous nodes in cooperative sensing for cognitive radio networks," Computers and Electrical Engineering, vol. 52, p. 284–292, 2016.

[11] U. S. Premarathne, I. Khalil and M. Atiquzzaman, "Trust based reliable transmissions strategies for smart home energy consumption management in cognitive radio based smart grid," Ad Hoc Networks, vol. 41, pp. 15-29, 2016.

[12] Y. Richard, T. Helen, H. Minyi, L. Zhiqiang and M. Peter, "Defense against Spectrum Sensing Data Falsification Attacks in Mobile Ad Hoc Networks with Cognitive Radios," 2009.

[13] W. Ren, R. Beard and E. Atkins, "A survey of consensus problems in multi-a gent coordination," in in Proc. American Control Conference 05, Portland, OR, June 2005.

[14] J. Wu, T. Song, Y. Yu, C. Wang and J. Hu, "Sequential cooperative spectrum sensing in the presence of dynamic Byzantine attack for mobile networks," PLoS ONE, vol. 13, no. 7, 2018.

[15] I. Ngomane, M. Velempini and V. S. Dlamini, "Trust-based System to Defend Against the Spectrum Sensing Data Falsification Attack in Cognitive Radio Ad Hoc Network," in International Conference on Advances in Big Data, Computing and Data Communication Systems, Durban, 2018.

Real-time Multiple Vehicle Detection using a Rear Camera Mounted on a Vehicle

Olivier Oheka and Chunling Tu
Department of Computer Systems Engineering
Tshwane University of Technology
Pretoria, South Africa
olivieroheka@gmail.com, tclchunling@gmail.com

***Abstract*—An increase in demand for traffic monitoring of densely populated areas has been the subject of many discussions concerning the security on the road. Cameras and other sensors are used to monitor, identify and manage potential incidents that can occur. Therefore, there is a need for computer vision system and algorithms to detect and track vehicles, which will facilitate the management of traffic and driving assistance. In this paper, a vehicle detection and tracking system is proposed using an efficient background cancelation technique and Haar-like features with a modified Adaboost algorithm in a cascade configuration for maximum accuracy and robustness. The proposed system is implemented on a passenger car with a camera mounted at the rear to detect vehicles behind it. Video data are collected and processed in real-time and the performance of the system is verified experimentally.**

Keywords—computer vision, vehicle detection, vehicle tracking, video surveillance, driving assistance

I. INTRODUCTION

Many aspects of our lives have become more enjoyable due to advanced technologies. The transport sector has taken quite a long time to catch up. Vehicles are becoming more and more secure; however, the road environment has become more complex, mainly due to the rapid increase in the number of vehicles and the resulting consequences such as traffic accidents, traffic congestion roads, hijacking etc. Cameras are commonly available on road facilities and vehicles, which provide huge potential for the development of traffic monitoring and vehicle tracking applications. The development of such computer vision systems can control and forecasting road traffic which is one of the priorities of intelligent traffic system (ITS) [1]. A driving assistant system is another important application helpful on increase the driving behavior monitoring and security [2]. These technologies aim to increase the safety, efficiency and convenience of traffic management and driving assistance, are currently being designed and developed [3].

This paper aims to construct a computer vision-based system, with rear looking cameras mounted on the host vehicle. The scene at the back of the host vehicle is captured and analyzed to detect and track other vehicles. This system will trigger alarm signals when potential hazards of incidents, such as the crash and over speeding. The system is applied to real traffic scene to validate the performance. The experiment results show the proposed system is effective.

II. RELATED WORK

During the last decade, numerous researchers have proposed methods of detecting and tracking vehicle movements, which is to separate vehicles from their environments in order to extract information useful for tracking processing, such as vehicle position, size, etc[4][5]. The tracking of moving vehicles in a sequence of images is currently a very active research topic in computer vision, the aim is to recognize vehicles from one background and to reconstruct their trajectory to extract the Information that characterizes each of them such as the distance traveled, the speed, etc.

Jin-Cyuan and Shih [6] have built a highway traffic monitoring system that can estimate traffic parameters such as the number and classes of vehicles, this system firstly extracts of the vehicle regions, and the detect continuation of movement, at last extract of the traffic parameters. The bottom subtraction method is first used to detect vehicles on the road. The two authors used some geometric properties to eliminate false detections and a shadow subtraction algorithm to improve detection accuracy. The tracking method used consists in associating the graphs (contours) to find the correspondence between the detected vehicles at different times. In this work the authors used videos with different lighting conditions to prove the effectiveness of their system.

Ehsan Adeli Mosabbeb et al. [7] used some static parameters for each frame of the video such as mean and standard deviation to detect and segment vehicles; This approach has been used for the detection of moving vehicles in a congested scene where it is necessary to find solutions to overcome the problem of false detections. In their work they proposed to subtract the shadow to better separate vehicles very close to each other. To show the robustness and precision of this approach they used road scenes under different climatic conditions such as noise, rain, snow, and varying lighting conditions.

Zhe Liu and Yangzhou Chen [8] proposed a system for estimating road traffic parameters based on motion detection by the consecutive frame difference algorithm, and the continuation by application of the camshaft algorithm Proven for tracking vehicles of varying shapes and sizes and under different light conditions. They then reconstructed the individual trajectory of each vehicle, giving the ability to estimate the individual parameters of each vehicle, such as distance traveled and speed.

François Bardet and Thierry Château [9] were interested in pursuing multiple objects in a road environment using one or more cameras, for this they used a method that combines a particulate filter with a Markov chain.

Xie Lei and Guangxi [10] used the Kalman filter for tracking vehicles in real time motion in a road scene, their work includes two steps: motion detection and tracking. They tested their algorithm on different scenes and for different climatic conditions. The results obtained proved the robustness of the algorithm in real time and for different light conditions.

The issue with most of these algorithms is that they focus too much on the difference between static background and moving foreground. They are weak against complex background, as Several inherent characteristics of natural outdoor environmental monitoring pose a challenge to these automated background modeling and subtraction. Namely, foreground objects tend to, by necessity, blend into the background, and the background exhibits large variations due to non-stationary objects (pedestrians, moving leaves) and sudden transitions from light to shadow. Also, most algorithms can accurately detect only a certain view of the object of interest, mostly due to the angle at which the video is taken. and accuracy decreases when the system is run attempted in well in real time because of all the computing process. They also tend to produce an erroneous result due to slow moving objects, lighting condition and many other challenges if introduced with.

These conditions present a challenge to the state of the art, which we have addressed with a simple but efficient algorithm that exhibits comparable performance also on standard surveillance data sets.

III. Proposed Vehicle Detection and Tracking System

A. System structure

The proposed flow chart for the system as shown in Figure 1, with the following steps.

1) Data Preparation

The data is captured from the camera and converted in a suitable form for initial processing.

2) Image Preprocessing

We prepare the input of our algorithm by separating the video into frames and filter the noise from the images to minimize false positives.

3) Image segmentation

We isolate the object of interest from the rest of the objects on the video by determining where the object of interest is likely to be detected.

4) Vehicle detection

With the use of the datasets containing positive and negative samples, the algorithm is compiled using Haar features, Adaboost and Cascade Amplifier technique and tested in real life conditions.

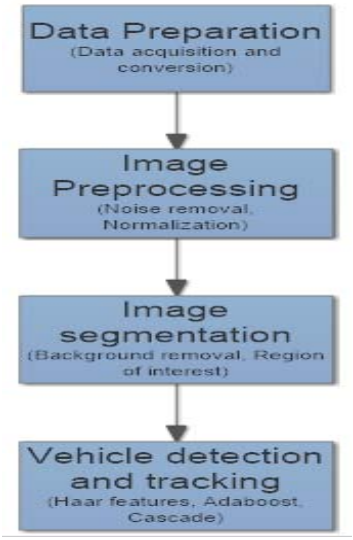


Figure 1. Structure of the proposed system

B. Data collection

A camera is placed at the rear of the vehicle as shown in Figure 2, in a way that it has a wide view to cover the two boundaries of the road. Then the video is acquired through serial port of the computer and stored in a video.

Figure 2. Vehicle with rear-mounted camera

C. Data conversion

A video is a rapid succession of still images (frames). Video systems vary in display resolution, aspect ratio, refresh rate, color capabilities and other qualities. For this research, towards the purposed of real-time vehicle detection, RGB color videos are considered with the frame rate of 25 fps and the resolution 620x340. The video will be processed in terms of frames.

D. Image preprocessing

1) Histogram Normalization

Brightness or darkness appeared sometimes in captured frame and can have bad impacts on the results. The RGB image is converted to the gray scale image for reduction of processed data, which also enhances the contrast in the spatial domain.

2) Noise Filtering

Noise existing in images, is any degradation in an image signal, caused by external disturbance while an image is being sent from one place to another place via satellite, wireless and network cable. In our case, the noise is presented in the input image when captured from the camera. There are many techniques for noise removal, such as low pass filtering in the

frequency domain, which may be a good choice, but this also removes some important information in the image. In our system median filtering is used for noise removal in the histogram normalized image for its simplicity and efficiency.

E. Background removal

A good and reliable visual object detection system generally starts the process by removing disturbance objects that are not of interest.

Background subtraction [11] is one of the most popular computer vision methods for object detection methods. Background subtraction methods build models for the background and then apply a decision function on a new frame to decide if points or region of points belong to the background or a moving object (foreground).

Given an image at the moment *t*, the simplest model is to consider the image at time *t - 1* as the background, and areas changing in appearance between *t - 1* and *t* are moving objects. In other words, motion detection is obtained by temporal derivation of the image sequence.

Many techniques [12] were developed based on this simple method for a static background, because it is capable to separate the foreground and background when the images contain the same and static background, such as the video captured for a road or a room by a statically mounted camera.

However, the method is not feasible for the moving imaging devices, which is common for vehicle-based systems. For instance, Figure 3 demonstrates a simple background or foreground detector cannot be used for the case of a moving background.

(a) Original frame (b) Current frame

(c) Previous frame (d) Foreground detection result

Figure 3. The problem of simple background subtraction

Figure 3 (a) and (b) represent the current and previous frame respectively. Traditional background subtraction technique makes use of the change of pixels position of different objects on the picture, meaning it removes the image areas having smaller position changes. But Figure 3(d) shows most pixels changed positions since the previous frame. And this situation is created because the optical flow of the camera changes with each frame.

In this paper, background removal starts from cutting the images and only focus on the area where a vehicle can appear. This step removes lots of background as shown in Figure 5. Then for the reminder of images, the component of background such as the road, view of trees, road lanes, and so on (as shown in Figure 4), as negative samples in classification algorithm (discussed later in the paper), to remove the background as a specific class. This allows to save time on the process of the algorithm.

With this classification method to remove background, the effects of the changes of the whole landscape will be minimized. Furthermore, the object of interest dominates with its size in the reminder region after cutting. This will make detection even easier as it will take less time in the algorithm computation for full background removal techniques and to detect the object of interest thanks to its bigger size on the segmented area.

Figure 4. Negative samples

F. Region of interest (ROI)

The searching area in the image is restricted by using common knowledge of the perspective of the road (Figure 4). As we know, vehicle cannot be found in the sky or in a building, they are found on the road. To confirm this theory, we capture footage of the road for a long period of time and use it to test our primary algorithm. The results showed that there is only a region on the image where vehicles can be found. The tracking algorithms were only performed on this cropped image region to reduce the processing time of the system.

The image is then examined by a sliding window evaluated by the detector at different resolutions, multiples of 32x32 the smallest acceptable resolution of a detected vehicle.

The vehicles are detected and tracked only when it enters the ROI and when the vehicles leave the ROI line the vehicles are stopped tracking. ROI is tracked by creating bounding boxes around the vehicles and the process are carried on. The ROI creation is as shown in the Figure 5.

Figure 5. Region of interest from original frame

Our region of interest, area inside the yellow rectangle box, in this case is found between y1=100 and y2 = 300. Considering this region of the frame gives the maximum result and is in perfect resonance with the angle of the rear-camera. The image size is 620x340.

G. *Vehicle detection*

There are many methods to detect objects on an image. Background subtraction, which is the most used, was discussed proved inefficient in our application. Other methods used are active contour, region and template-matching based method. Above approaches can effectively accomplish detection and tracking tasks, but they need more system computation and have certain application conditions. So, to reduce calculation time we use learning-based approaches to detect and track video vehicles efficiently to improve overall system accuracy and robustness.

The system makes use of an Adaboost cascade classifier with Haar-like features. Haar features was proposed to use characteristics, a synthetic representation and informative, calculated from pixel values [13]. A simple method for the Haar features calculation is to use the difference of the sums of pixels from two or more adjacent rectangular areas. A set of vehicles images are used to extract these features, as demonstrated in the Figure 5.

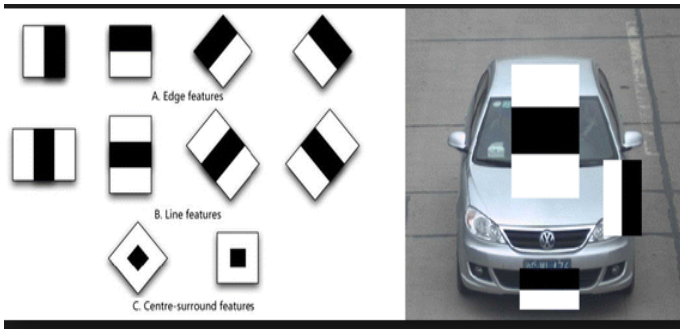


Figure 5. Haar features of a vehicle

AdaBoost is a type of "Ensemble Learning" where multiple learners are employed to build a stronger learning algorithm. AdaBoost works by choosing a base algorithm and iteratively improving it by accounting for the incorrectly classified examples in the training set.

Equal initial weights are assigned for all the training examples and a base algorithm is chosen to train weak learners, which consequently increases the weights of the incorrectly classified samples. The training is iterated by *n* times, each time applying base learner on the training set with updated weights.

The final decision is made by the weighted sum of the *n* learners [14].

To implement this ensemble learning strategy, multiple classifiers are combined in a cascade manner [15], where each is based on the results of one or more different Haar features. The cascade classifier consists of a list of stages, where each stage consists of a list of weak learners. The algorithm detects the vehicle by placing and moving a sliding window over the image. Each stage of the classifier labels the specific region defined by the current location of the sliding window as either positive (vehicle detected) or negative (no vehicle found).

If the labelling yields a negative result, then the classification of this specific region is hereby complete, and the location of the sliding window is moved to the next location.

If the labelling gives a positive result, then the region (multiples of 32x32) moves to the next stage of classification. When the object passes through all stages of classifiers, it gets recognized. In this paper, three stages for maximum are considered.

To further increase accuracy of our system, we use the initial footage recorded by the camera for preliminary tests to separate detected vehicles images to feed them back in the positive sample database and remaining background in the negative sample database. This modified approach increases accuracy of the system

H. *Vehicle tracking*

Vehicle tracking is the process to detect vehicles on consecutive frames, as illustrated in figure 6. The camera gives data at a certain frame rate and the processing time of our detection algorithm must be faster than the used frame rate to ensure robustness of the system.

Figure 6. Tracking process

IV. RESULTS AND ANALYSIS

All images were captured by a camera located inside of the host vehicle, for urban and rural areas. 560 images/frames of traffic scene were used as positive images and 1000 were used for negative samples. Vehicles appear in the images contains various models, such as passenger cars, bus and trucks. Each detected vehicle is labeled as “car” and is framed by a rectangle box. The positive sample set is doubled by synthesizing a mirror image by axial symmetry. Of the total 1120 images, two-thirds are used for the positive learning set and the rest for the positive validation set. The latter serves to adjust the threshold *S* of the classifier based on a rate of correct detections and the rate of false alarms.

Some frames with the detected vehicles are shown in Figure 7 as examples.

Table I shows the system is effective for different categories of vehicles. Of course, this depends on the number of positive and negative images used for the training algorithm.

Figure 8 shows the vehicle detection success rate of the system in terms of true positive rate, true negative rate, false positive rate and false negative rate. The proposed got high success rate on both true positive and true negative, which means high reliability is obtained. The overall computation

time of the system to detect all vehicles on a frame is 51ms, resulting in the system to work perfectly with a frame rate up to 20 frames per second, which is really good.

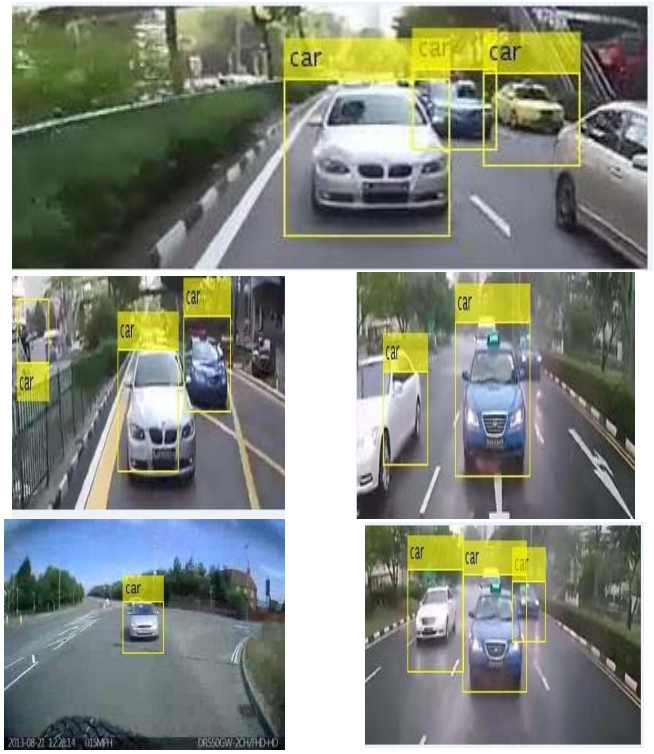


Figure 7 : Detected vehicles

TABLE I. SUCCESS RATE FOR DIFFERENT TYPES OF VEHICLE(TRUE POSITIVE)

Vehicle types	Method in [6]	Method in [7]	Method in [8]	Method in [9]	Proposed method
Car	87%	77%	89%	93%	**97%**
Truck	80%	76%	80%	**92%**	**92%**
Bus	50%	65%	79%	78%	**93%**

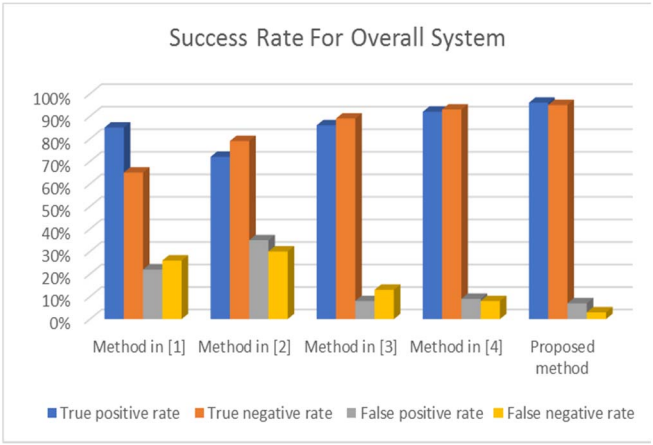


Figure 8. Success rate for overall system compared with existing methods

V. CONCLUSION

In this paper we presented a consistent on-road multiple vehicles detection method from video sequences. Data were collected through a camera mounted at the back of the host vehicle. The vehicles in traffic scene were effectively detected, counted, and tracked, and classified during different weather conditions. The results obtained are satisfactory in the sense of low false positives and high true positives. The main contribution of the system is an efficient vehicles detection and tracking in real and different environments with a moving camera by using a simple but efficient background removal technique and a modified Adaboost algorithm. Small issues have been faced in the experiment, such as lighting conditions, vehicles shapes distinctions and real time situations and need to be addressed in further researches to improve the overall system performance to a higher level.

ACKNOWLEDGMENT

This work is based on the research supported by the National Research Foundation of South Africa (Grant Number 99383).

REFERENCES

[1] R. A. Hadi, G. Sulong and L. E. George, "Vehicle detection and tracking techniques: A concise review", Signal & Image Processing: An International Journal (SIPIJ), vol. 5, No.1, February 2014.

[2] B. S. Mokha and S. Kumar, "A review of computer vision system for the vehicle identification and classification from online and offline videos," Signal & Image Processing: An International Journal(SIPIJ), vol. 6, No.5, October 2015.

[3] M. Chaple, S.S. Paygude, V. Vyas, "Vehicle detection and tracking using optical flow and background subtraction technique," International Conference on advances in Computer Science and Application, Lucknow , June 21-22, 2013.

[4] M. Chaple and S.S. Paygude, "Object detection and tracking in video frame sequence," National Conference: Institute for Research & Development India, Nagpur, 6th Nov 2012.

[5] M.Teutsch and W.Kruger, "Robust and fast detection of moving vehicles in aerial videos using sliding windows," The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2015

[6] J. Lai and S. Huang, "Image-based vehicle tracking and classification on the highway," The 2010 International Conference on Green Circuits and Systems, Shangai, China, 21-23 June 2010

[7] E. A. Mosabbeb, M. Sadeghi and M. Fathy, "A new approach for vehicle detection in congested traffic scenes based on strong shadow segmentation," ISVC 2007, Part II, LNCS 4842, pp. 427–436.

[8] Z. Liu, Y. Chen and Z. Li, "Camshift-based real-time multiple vehicles tracking for visual traffic surveillance," World Congress on Computer Science and Information Engineering, 2009, Los Angeles, CA, USA, 31 March-2 April 2009.

[9] F. Bardet and T. Château, "MCMC particle filter for real-time visual tracking of vehicles," Proceedings of the 11th International IEEE, Conference on Intelligent Transportation Systems, Beijing, China, pp 539 - 544, 2008.

[10] L. Xie, G. Zhu and Z. Zhang, "Real-time vehicles tracking based on kalman filter in a video-based ITS," Proceedings of the 2005 International Conference on Communications, Circuits and Systems Hong Kong, China, 27-30 May 2005.

[11] K. Yang, Z. Cai and L. Zhao., "Algorithm Research on Moving Object Detection of Surveillance Video Sequence," Optics and Photonics Journal 2013, vol. 3, pp 308-312.

[12] P. Syroka and R. Hudec, "Comparison of background subtraction methods on near infra-red spectrum video sequences," Procedia Engineering, vol. 192, pp 887-892.

[13] Y. Tang1, Y. C. Xu1 and C. Z. Zhang, "Robust vehicle detection based on cascade classifier in traffic surveillance system," The Open Automation and Control Systems Journal, 2014, vol. 6, pp 349-354.

[14] R. Lienhart and J. Maydt, "An extended set of Haar-like features for rapid object detection," In IEEE International Conference on Image Processing. USA pp. 900-903, 2002.

[15] P. Viola, and M. Jones, "Rapid object detection using a boosted cascade of simple features," Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. USA, pp. 511-518, 2001.

Artificial Neural Networks in Water Distribution Systems: A Literature Synopsis

Thapelo Mosetlhe [1,2,*], Yskandar Hamam [1,2,3], Shengzhi Du [1,2] and Yasser Alayli [4]
[1] Tshwane University of Technology, Department of Electrical Engineering, Staatsartillerie Rd, Pretoria, 0183, South Africa
[2] French South African Institute of Technology (FSATI), Pretoria, 0001, South Africa
[3] École Supérieure d'Ingénieurs en Électrotechnique et Électronique, 2 Boulevard Blaise Pascal, 93160 Noisy-Le-Grand, France
[4] Laboratoire d'Ingénierie des Systémes de Versailles, UVSQ, Université Paris-Saclay, 10-12 Avenue de l'Europe, Vélizy, 78140
[*] Email: mosetlhethapelo@yahoo.com; Tel: +27 71 185 4301

***Abstract*—High computational requirements are commonly associated with the hydraulic simulation of large-scale water distribution. The convergence of the cumbersome iterative procedures involved has been a well-debated issue for the past decades. The large-scale and non-linear properties pose a great hindrance towards the development of online applications for water distribution network (WDN) analysis and pressure control thereof. Consequently, there has been a great interest in the deployment of model-free techniques to mimic the rather computationally expensive non-linear hydraulic simulations. As the hydraulic simulation based research is still being conducted, the advantages of model-free techniques make them more suitable alternatives. Artificial neural networks (ANN) is one of the most successful model-free methods for WDN analysis and management. In this paper, a literature synopsis of existing applications of model-free approaches in water distribution is presented. The technical advantages of applying such technique in a large-scale non-linear network are brought up in this paper.**

***Keywords*—Hydraulic simulation, Water Distribution Network, Convergence, Pressure control, Model-free control, Artificial neural networks.**

I. Introduction

The ever and steady increase of population is leading to the stressing the availability of water resources. The water transportation systems are also burdened by this phenomenon. The water resource transportation system protection and reliability form one of the most critical elements of day to day responsibilities of the water supply utilities or municipalities. To meet the ever-increasing and stochastic demands, the operating conditions of the water distribution network (WDN) is continuously adjusted. This makes the timing crucial as miscalculations or delays may be catastrophic to the network and the supply. The computation associated with the hydraulic simulation hinders the progress towards real-time analysis and adjustment. However, efforts towards achieving the real-time water distribution network are geared up [1], [2]. Supervisory control and data acquisition (SCADA) systems are commonly deployed to transmit the measurements to avoid educated guesses of demand [3] within the hydraulic simulation [1], [4], [5]. Recently, to minimise the computations, artificial neural networks (ANN) are replacing the time consuming hydraulic non-linear simulations [5], [6].

The advantage of ANN in WDN is reported by *Hamam* and *Hindi* [7]. However, to effectively model the hydraulic behaviour of the water distribution network, a large number of simulation has to be carried out to generate the training data [6]. An insufficiently train ANN can render this replica ineffective.

Therefore, to apply an ANN, the hydraulic model must be defined and solutions be obtained for various inputs. The harmonised hydraulic demand-driven model is widely available in the literature [8], [9]. However, these solutions have been an issue well debated in the past decades [10], [11], [12], [13], [9], [14], [15]. The common denominator for all the methods includes heavy iterative procedures and distinguished convergence properties. The global gradient algorithm (GGA) [9] has been widely used due to easier implementation and it is also used in readily available package EPANET 2 [16].

The computational requirement is the driving force towards model-free techniques which can be suitable for real-time applications. This paper reviews existing literature on model-free techniques applied to water distribution networks. As the research in this field is still in the early stage of development, this work provides synopsis which subsequent investigation may draw benefit.

This paper is organised as follows: Section II presents the classical mathematical model for the steady state analysis of the WDN. III presents the replication of the non-linear hydraulic simulation using ANN with challenges and advantages of such replications presented. The ANN applications in network and pressure management of WDN techniques are presented in Section IV. Leakage estimation and simulation in water distribution networks are presented in Section V. Section VI presents efforts to asses the water quality. The conclusion is furnished in the last section.

II. Mathematical Model of Water Distribution Network

For readers convenience, we recall the rather harmonised hydraulic equations for water distribution shown in Figure 1. The nodal balance of WDN consisting of N_b pipes and N number of nodes is given by [15], [17]

$$\sum CQ = I \tag{1}$$

where, $I \in \forall \quad N$ is a vector of nodal injection and demands and $Q \in \forall \quad N_b$ is a vector of flows across all the pipes. C_{ij}

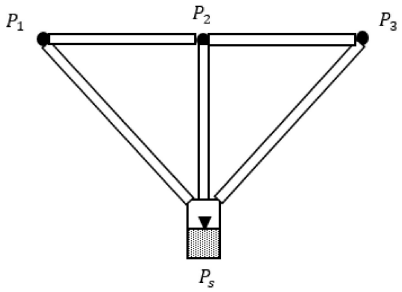


Fig. 1. Example of water distribution network

in $N \times N_b$ node-branch incidence matrix and

$$C_{ij} = \begin{cases} +1, & \text{if flow in branch } j \text{ leaves node } i \\ -1, & \text{if flow in branch } j \text{ enters node } i \\ 0, & \text{if branch } j \text{ is not incident to node } i \end{cases} \quad (2)$$

We decompose C to form

$$C = \begin{bmatrix} C_s \\ C_l \end{bmatrix} \quad (3)$$

and for load nodes, the nodal balance equation can be written as

$$C_l Q = -D \quad (4)$$

For the pipe l shown in Figure 2, the pressure drop across my be computed using Equation 5

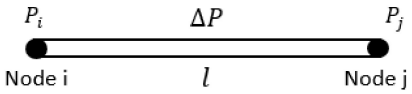


Fig. 2. A segment of water distribution network

$$\Delta P = P_i - P_j = k_p Q_l^{\alpha} = k_p Q_l |Q_l|^{\alpha - 1} \quad (5)$$

where k_p is the pipe hydraulic resistance. The matrix form of the energy conservation of the pipes may represented as

$$\Delta P = \begin{bmatrix} C_s^T & C_l^T \end{bmatrix} \begin{bmatrix} P_s \\ P_l \end{bmatrix} \quad (6)$$

where, P_s and P_l are vectors of pressures at sources and load nodes respectively. Equations 5 and 6 yields

$$k_p Q_l |Q_l|^{\alpha - 1} - C_s^T P_s - C_l^T P_l = 0 \quad (7)$$

if matrix A is defined as

$$A = diag\left(k_p |Q|^{\alpha - 1}\right) \quad (8)$$

then the system of hydraulic equations may be written as

$$\begin{aligned} AQ - C_s^T P_s - C_l^T P_l = 0 \\ C_l Q + D = 0 \end{aligned} \quad (9)$$

III. ANN in Modelling Water Distribution Network (WDN)

The mathematical model of the water distribution network is widely available in the literature as systems of equations representing the nodal balance and the energy conservation of the pipes. Such equations are recalled in Section II. As the solutions involve iterative procedures, excessive computation is required to solve them. When embedded in an optimisation problem it becomes a major challenge hindering the progress in the real-time management of WDN [18], [5]. Therefore, the efficient approaches to avoid these iterative procedures when modelling input/output relationship is currently the topic being investigated widely.

ANN is a common tool for modelling complex non-linear systems. Consequently, in the early 90s, *Hamam* and *Hindi* [7] applied ANN to map the hydraulic behaviour of non-linear WDN with the backpropagation used as the training algorithm. The structure of ANN used in [7] is shown in Figure 3 and for the inputs x_i and the weight ω_i, the output of neuron i can be expressed as (10) [7], [18]

$$y_i = f\left(\sum \omega_i x_i\right) \quad (10)$$

where

$$f(y_i) = \frac{1}{(1 + e^{-a y_i})} \quad (11)$$

and a is a constant used to control the slope of semi-linear region [19]. The suitability of ANNs in real-time water distribution is recognised by *Rao* and *Salomons* in [5]. Potable Water Distribution Management (POWADIMA) research project proposed multi-layer, feed-forward to mimic the operation of WDN [5], [6], [18]. More recent, flow dynamics of the water distribution network were forecasted using multi-layer feed-forward back-propagation ANN with up to 99 % confidence [20]. It is worth noting that for successful replications, the large number of Off-line simulations are with different inputs combinations are required [6].

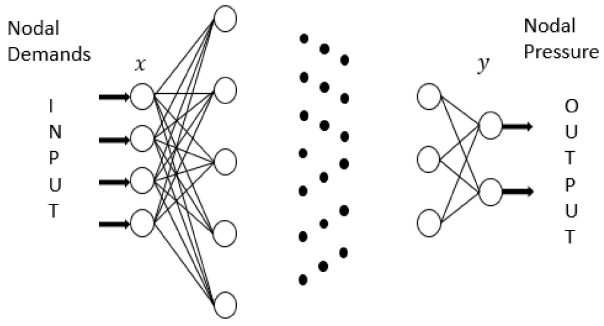


Fig. 3. Structure of ANN with multiple hidden layers

IV. WDN and Pressure Management

Pressure management is one of the most important aspects of network operation. Proper management meets the demand from WDN adequate pressures. Furthermore, an adequate pressure in the network results in optimal operation of the WDN and therefore lengthening the lifespan of the infrastructure. Due to these important considerations, an artificial neural network (ANN) was used to evaluate the pressure heads and

pressure zones in [21]. The essence of pressure management in their work is to determine pressure zones and locating the areas for the installation of pressure reducing stations.
Limitations in resources to monitor the pressure in the WDN is a major challenge towards effective pressure management. Therefore, it is often impossible to measure all nodes of the network. As a result, available measurements are used to estimate the unavailable data. To achieve this, ANN is used to estimate the pressure profile of the WDN based on the available data [22]. The estimated data was then validated by comparing with the results of normal hydraulic simulations.

V. ANN in Leakage Analysis and Minimisation

Leakage in WDN is an issue of great concern to the municipalities and water utilities as it contributes to the exhaustion of water resource [23]. Therefore, it is imperative that they are continuously detected and localised [8] for speedy interventions. Various models are developed in the literature to estimate the leakages in the network [23], [24], [25]. From the results of the models, ANNs are trained to estimate the leakage in the water distribution system [26]. The strategy was base on single layer as shown in Figure 4 and multi-layer feed-forward as depicted by Figure 3. The ANN estimated leakage is compared to principal component analysis with multi-layer scheme providing better accuracy. Furthermore, *Makaya* and *Hensel* [20] developed a leakage detection technique based on ANN forecasted flow dynamics.

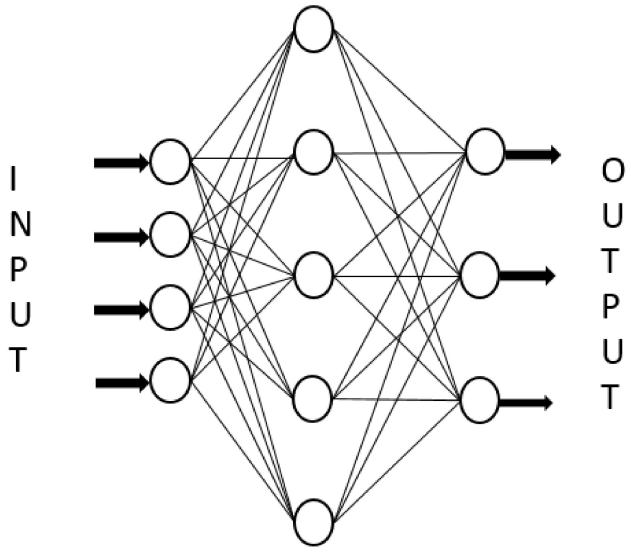


Fig. 4. Structure of ANN with single hidden layer

VI. ANN in Water Quality Assessment

Water quality is one of the most important aspects involved in water network management. It is required by regulations that water delivered to the consumers must meet certain quality requirements to ensure that their health is not compromised. In the work of *Adedoja et al.* [27], various water quality models are interrogated extensively. The advantages and disadvantage are presented and suggestion to embed hydraulic analysis in water quality model and contamination source identification is made. In [28], historic parameters known to influence the quality of water are used to train the ANN model. The developed model as shown in Figure 5 is used to estimate the quality of water. Similar approach, with difference being the

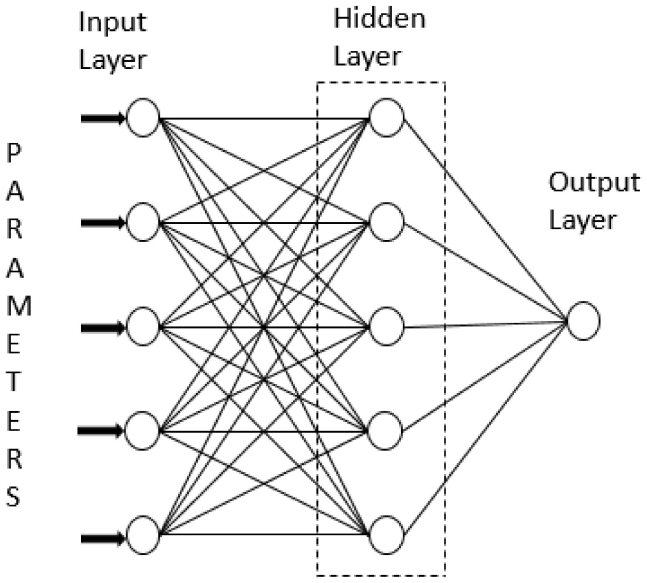


Fig. 5. ANN structure for water quality assessment [28]

input parameters was adopted in [29]. The dissolved oxygen where computed together with the biochemical oxygen demand as the measure of the water quality [29].
In [30], data obtained from remotely located sensors is fed to a previously defined ANN model to estimate the water quality.

VII. Other Applications of ANN in WDN

A. Demand Forecasting

Demand forecasting in WDN ensures that the network is optimally operated to meet the requirements. As the demand is time-varying and stochastic, the prediction of its uncertainty becomes an important aspect of network operation. Various methodologies for time-series data modelling were compared in [31] and model conditional processor was adopted in [32]. The efficiency of ANNs pioneered the ditching of conventional regression and time series towards knowledge-based prediction schemes in water demand forecasting [33], [34]. In [35] the conventional methods are compared to ANNs and wavelet artificial neural networks (WA-ANN). *Ghiassi et al.* used dynamic artificial neural networks (DANN) model with 99 % accuracy for monthly, weekly and daily data [36]. This ANNs has been used for short-term water demand forecast [37], peak daily water demand forecast [38] amongst other application due to their efficiency as compared to regression models.

VIII. Summary of ANN Applications in WDN

Table I presents the summary of some existing approaches furnished by researchers in order to address the ANN application in water distribution networks.

TABLE I. APPLICATIONS OF ANN IN WDN

Applications	Remarks	ANN Type
ANN in Modelling WDN	Large number of simulation are required. However, the overall computation is reduced and possibility of real-time application is viable.	Back-propagation [7], Multi-Layer Feed-Forward back-propagation [20], Three-layer Feed-Forward [18]
ANN in WDN and Pressure Management	Knowledge based trained controller eliminate the iterative optimisation procedure to optimally manage the pressure in WDN.	Back-propagation [7], Three-layer [22]
ANN in Leakage analysis and minimisation	A trained model can be used for any network since as long as the parameters of the leakage are the same as those used in the learning phase.	Multi-layer feed-forward back-propagation [20], back-propagation [26]
ANN in Water quality modelling	A trained model can be used for any network since as long as the parameters of the water quality are the same as those used in the learning phase.	Feed-forward back-propagation [28], feed-forward, fully-connected, three-layer perceptron neural network [39]
Other applications of ANN in WDN	ANN trained model can predict demand more accurately the conventional time-series and regression algorithms	Back-propagation ANN [33], feed forward two-layer network [34], Three-layer feed-forward ANN [35]

IX. DISCUSSION AND CONCLUSION

Generally, cumbersome computations are performed to find solutions of large-scale WDN. However, this burden is not feasible for real-time applications. The computation even increases when embedded in an optimisation problem. Therefore, the feasible solution is to employ model free replication of flow dynamics of the water distribution systems. Under this situation, ANN is widely used due to the high efficiency compared to methods based on hydraulic simulation, which enables efficient estimation of the flows and the unknown pressure heads in WDN. Based on ANN, more application such as pressure management, leakage management, water quality analysis can be achieved with far lesser computational demand. Therefore, by adopting the ANN to mimic the non-linear hydraulic analysis can be beneficial as it makes the online applications more feasible. With the advance in technologies and the recent introduction of deep learning algorithms, the role of the hydraulic simulation may be limited to a tool for the generation of datasets. It can be seen that efforts in various aspects of water distribution are geared up towards artificial intelligence and therefore, making the WDN modelling using ANN a topic of more of interest.

ACKNOWLEDGMENT

The authors would like to thank Tshwane University of Technology for providing financial support and resources to carry out the research.

REFERENCES

[1] D. Kang, "Real-time optimal control of water distribution systems," *Procedia Engineering*, vol. 70, pp. 917–923, 2014.

[2] N. Fontana, M. Giugni, L. Glielmo, G. Marini, and F. Verrilli, "A lab prototype of pressure control in water distribution networks," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 15 373–15 378, 2017.

[3] W. Cheng, T. Yu, and G. Xu, "Real-time model of a large-scale water distribution system," *Procedia Engineering*, vol. 89, pp. 457–466, 2014.

[4] E. Creaco and M. Franchini, "A new algorithm for real-time pressure control in water distribution networks," *Water Science and Technology: Water Supply*, vol. 13, no. 4, pp. 875–882, 2013.

[5] Z. Rao and E. Salomons, "Development of a real-time, near-optimal control process for water-distribution networks," *Journal of Hydroinformatics*, vol. 9, no. 1, pp. 25–37, 2007.

[6] D. G. Jamieson, U. Shamir, F. Martinez, and M. Franchini, "Conceptual design of a generic, real-time, near-optimal control system for water-distribution networks," *Journal of Hydroinformatics*, vol. 9, no. 1, pp. 3–14, 2007.

[7] Y. Hamam and K. Hindi, "Optimised on-line leakage minimisation in water piping networks using neural nets," in *Proceedings of the IFIP Working Conference, Dagschul, Germany*, vol. 28, 1992, pp. 57–64.

[8] K. Adedeji, Y. Hamam, B. Abe, and A. Abu-Mahfouz, "Leakage Detection and Estimation Algorithm for Loss Reduction in Water Piping Networks," *Water*, vol. 9, no. 10, p. 773, Oct. 2017. [Online]. Available: http://www.mdpi.com/2073-4441/9/10/773

[9] E. Todini and S. Pilati, "A gradient algorithm for the analysis of pipe networks," in *Computer applications in water supply: vol. 1—systems analysis and simulation*. Research Studies Press Ltd., 1988, pp. 1–20.

[10] Y. Hamam and A. Brameller, "Hybrid method for the solution of piping networks," in *Proceedings of the Institution of Electrical Engineers*, vol. 118. IET, 1971, pp. 1607–1612.

[11] M. Hall, "Hydraulic network analysis using (generalized) geometric programming," *Networks*, vol. 6, no. 2, pp. 105–130, 1976.

[12] M. Collins, L. Cooper, R. Helgason, J. Kennington, and L. LeBlanc, "Solving the pipe network analysis problem using optimization techniques," *Management science*, vol. 24, no. 7, pp. 747–760, 1978.

[13] P. Carpentier, G. Cohen, and Y. Hamam, "Water network equilibrium, variational formulation and comparison of numerical algorithms," *EURO VII*, 1985.

[14] Y. Hamam and K. Hindi, "Steady-state solution of pipe networks: a

new efficient optimisation-based algorithm," *DTG Report, Computation Department, UMIST*, 1989.

[15] Y. Hamam, "Steady state solution of water piping networks," 2017, unpublished report.

[16] L. A. Rossman, "Epanet 2 users manual, water supply and water resources division," *National Risk Management Research Laboratory, US Environmental Protection Agency, Cincinnati, Ohio*, 2000.

[17] L. Mpiana, Y. Hamam, and A. M. Abu-Mahfouz, "Graph based hydraulic modelling of pressure in water distribution networks," in *AFRICON, 2017 IEEE*. IEEE, 2017, pp. 1558–1563.

[18] Z. Rao and F. Alvarruiz, "Use of an artificial neural network to capture the domain knowledge of a conventional hydraulic simulation model," *Journal of Hydroinformatics*, vol. 9, no. 1, pp. 15–24, 2007.

[19] M. Huiqun and L. Ling, "Water quality assessment using artificial neural network," in *Computer Science and Software Engineering, 2008 International Conference on*, vol. 1. IEEE, 2008, pp. 13–15.

[20] E. Makaya and O. Hensel, "Modelling flow dynamics in water distribution networks using artificial neural networks-a leakage detection technique," *International Journal of Engineering, Science and Technology*, vol. 7, no. 1, pp. 33–43, 2015.

[21] J. Dawidowicz, "Evaluation of a pressure head and pressure zones in water distribution systems by artificial neural networks," *Neural Computing and Applications*, pp. 1–8, 2017.

[22] E. Ridolfi, F. Servili, R. Magini, F. Napolitano, F. Russo, and L. Alfonso, "Artificial neural networks and entropy-based methods to determine pressure distribution in water distribution systems," *Procedia Engineering*, vol. 89, pp. 648–655, 2014.

[23] L. S. Araujo, H. Ramos, and S. T. Coelho, "Pressure Control for Leakage Minimisation in Water Distribution Systems Management," *Water Resources Management*, vol. 20, no. 1, pp. 133–149, Feb. 2006. [Online]. Available: http://link.springer.com/10.1007/s11269-006-4635-3

[24] P. D. Dai and P. Li, "Optimal Localization of Pressure Reducing Valves in Water Distribution Systems by a Reformulation Approach," *Water Resources Management*, vol. 28, no. 10, pp. 3057–3074, Aug. 2014. [Online]. Available: http://link.springer.com/10.1007/s11269-014-0655-6

[25] O. Giustolisi, D. Savic, and Z. Kapelan, "Pressure-driven demand and leakage simulation for water distribution networks," *Journal of Hydraulic Engineering*, vol. 134, no. 5, pp. 626–635, 2008.

[26] D. Jang, H. Park, and G. Choi, "Estimation of leakage ratio using principal component analysis and artificial neural network in water distribution systems," *Sustainability*, vol. 10, no. 3, p. 750, 2018.

[27] O. Samson Adedoja, Y. Hamam, B. Khalaf, and R. Sadiku, "Towards development of an optimization model to identify contamination source in a water distribution network," *Water*, vol. 10, p. 579, 04 2018.

[28] G. C. Cordoba, L. Tuhovčák, and M. Tauš, "Using artificial neural network models to assess water quality in water distribution networks," *Procedia Engineering*, vol. 70, pp. 399–408, 2014.

[29] K. P. Singh, A. Basant, A. Malik, and G. Jain, "Artificial neural network modeling of the river water qualitya case study," *Ecological Modelling*, vol. 220, no. 6, pp. 888–895, 2009.

[30] Y. Chebud, G. M. Naja, R. G. Rivero, and A. M. Melesse, "Water quality monitoring using remote sensing and an artificial neural network," *Water, Air, & Soil Pollution*, vol. 223, no. 8, pp. 4875–4887, 2012.

[31] A. O. Anele, Y. Hamam, A. M. Abu-Mahfouz, and E. Todini, "Overview, comparative assessment and recommendations of forecasting models for short-term water demand prediction," *Water*, vol. 9, no. 11, p. 887, 2017.

[32] A. O. Anele, E. Todini, Y. Hamam, and A. M. Abu-Mahfouz, "Predictive uncertainty estimation in water demand forecasting using the model conditional processor," *Water*, vol. 10, no. 4, p. 475, 2018.

[33] J. Liu, H. H. Savenije, and J. Xu, "Forecast of water demand in weinan city in china using wdf-ann model," *Physics and Chemistry of the Earth, Parts A/B/C*, vol. 28, no. 4-5, pp. 219–224, 2003.

[34] I. S. Msiza, F. V. Nelwamondo, and T. Marwala, "Water demand prediction using artificial neural networks and support vector regression," 2008.

[35] J. Adamowski, H. Fung Chan, S. O. Prasher, B. Ozga-Zielinski, and A. Sliusarieva, "Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in montreal, canada," *Water Resources Research*, vol. 48, no. 1, 2012.

[36] M. Ghiassi, D. K. Zimbra, and H. Saidane, "Urban water demand forecasting with a dynamic artificial neural network model," *Journal of Water Resources Planning and Management*, vol. 134, no. 2, pp. 138–146, 2008.

[37] A. Jain, A. K. Varshney, and U. C. Joshi, "Short-term water demand forecast modelling at iit kanpur using artificial neural networks," *Water resources management*, vol. 15, no. 5, pp. 299–321, 2001.

[38] J. F. Adamowski, "Peak daily water demand forecast modeling using artificial neural networks," *Journal of Water Resources Planning and Management*, vol. 134, no. 2, pp. 119–128, 2008.

[39] N. M. Gazzaz, M. K. Yusoff, A. Z. Aris, H. Juahir, and M. F. Ramli, "Artificial neural network modeling of the water quality index for kinta river (malaysia) using water quality variables as predictors," *Marine pollution bulletin*, vol. 64, no. 11, pp. 2409–2420, 2012.

Fuzzy Logic Non-Linear Model Predictive Control of the Grinding Process in Ball Mills

Kiran Ninan
Department of Mechanical Engineering, Mechatronics, and Industrial Design
Tshwane University of Technology
Pretoria, South Africa

Shengzhi Du, BJ van Wyk
Department of Electrical Engineering
Tshwane University of Technology
Pretoria, South Africa

Chunling Tu
Department of Computer Systems Engineering
Tshwane University of Technology
Pretoria, South

***Abstract*—Grinding in ball mills is a crucial technological and industrial process which is used for the reduction of the size of particles with variant physical, chemical and mechanical characteristics. The control performance of the ball mills' grinding process is of outmost importance as this will determine the profit, where the energy consumption, the product quality and time efficiency are commonly concerned. In this paper, nonlinear model predictive control for ball mill grinding process is implemented. Economic performance, time delays and the consumption of energy in the grinding process with the proposed control system are engaged using Discrete Element Method (DEM) software. The results from experimental tests indicate the proposed method to be effective.**

Keywords—grinding; ball mills, control performance, energy consumption, DEM - Discrete Element Method

I. INTRODUCTION

Advanced Process Control (APC) is a common integral part in many industries, in increasing the profit margins, such as low energy consumption. Economic performance of such control systems is one of the important positives from upgrading projects.

Non-steady operations are a common factor in the process industrial fold and the solution to non-linear and dynamic problems are of high importance. Non-Linear Model Predictive Control (NMPC), Fuzzy Logic and Dynamic Real-Time Optimization (DRTO) are emerging industrial technologies and which is more viable and appropriate in the current industrial climate.

The application of ball mills is highly important in the processing of minerals and in the chemical industries.. The application reduce sizes of different sets of particles with different characteristics makes grinding in ball mills of very high importance with economic views. The efficiency and quality of the output from the ball mills is of outmost importance as this will determine the profit, since the process is closely relevant to the quality of the production.

The primary objective of the paper is to incorporate process control analysis into the grinding in ball mills. Economic performance analysis, time delays and the consumption of energy are analyzed and studied for further improvement.

Most of the current ball mills in use, do not take into consideration or sometimes lack aspects such as system nonlinearity, uncertainties, economic efficiency and production quality and dynamics.

Ball mill grinding process control encounters many disturbances or characteristics such as technological constraints, time variations of the various parameters, which are non-linear and not explicitly measurable.

The research aims to find a method to enhance the control performance of the process of grinding in ball mills and increase profitability of the mining companies by improving the economic performance of the process and implementing methods to improve the process as a whole.

II. RELATED WORK

2.1 FUZZY LOGIC

Fuzzy Logic is a way of computing method based on a truth of a certain degree other than actual definite values or answers.

To understand the concept of fuzzy logic, it must be known that it has a couple of meanings. Firstly, fuzzy logic is a straight forward, logical system and has an extension of logic of many values. Secondly, in a broader sense, the theory of fuzzy sets, relates to arbitrary boundaries where the membership of classes of objects is of a degree. Fuzzy logic is strongly based on precision and the importance of it. Where the question arises of how important is it to be exactly right, or will the rough answer do?

Taking this into consideration that fuzzy logic is primarily old and also quite new in terms of modern methodology, but the basic concept of this logic is derived from age old skills from years back [1]. The starting point of any control system is to map the inputs to the outputs and this is why fuzzy logic is so convenient and easy to map the input space and output.

A controller that uses a set of rules to indicate or specify values with the aid of fuzzy logic rules is regarded as a fuzzy logic controller. [2]. In a general view, it is regulating system and the measurements are computed into a fuzzified form and the end result is defuzzified, hence, a specific number is transformed back.

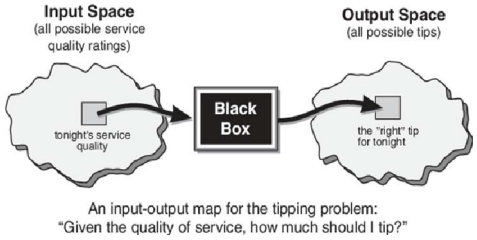


Figure 1: Fuzzy Logic Mapping

2.2 NON -LINEAR MODEL PREDICTIVE CONTROL

Non-Linear Model Predictive Control (NMPC) can be used on a system as a computational analysis which can be repeated on the system with state constraints, dynamics of the system and a finite horizon open-loop control problem.

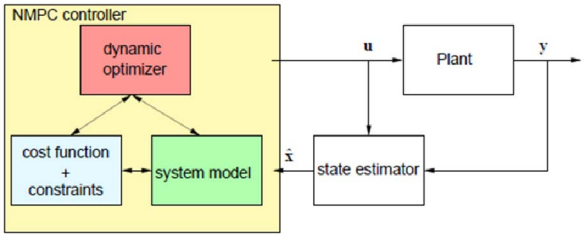


Figure 2: Control Loop of a Non-Linear Model Predictive Process

There are a few characteristics of Non-linear Model Predictive Control (NMPC), the few that make NMPC unique are - nonlinear model is allowed for prediction; input and state constraints is considered explicitly; closed loop behaviour is generally different from a predicated behaviour; prediction of the system in question needs to be estimated or measured first before any further assumptions or predictions [3].

According to the principle of optimality the open loop predictions will match up with the closed loop trajectories. In infinite horizon schemes, the main issue will often and usually not be applied in practical cases, since the open loop control problem cannot be optimally solved efficiently fast enough [4].

The frequently used approach for Nonlinear Model Predictive Control output-feedback is based on the certainty equivalent principle [6]. The estimate state x is measured with a stable observer and this is used in the model predictive controller. If one assumes that the observer error is exponentially stable, then only or most frequently only local stability is achieved of a closed loop process. Hence, the error entailed by the observer needs to be minimal in order to guarantee stability for a closed loop process. There is an issue that exists for non-linear systems that there is no general, across-the-board valid separation principle, that being said, the approach mentioned is implemented successfully in applications [7].

NMPC problems should also be formulated efficiently for a good outcome. Different possibilities are viable for an NMPC which are - without losing stability or performance by using short horizon length [8]; Use of suboptimal Nonlinear Model Predictive Control strategies.

Some cases, systems with a valid control Lyapunov function, feedback linear systems or stable systems can be removed. Such a method gives the possibility for an NMPC scheme that is efficient enough to formulate computationally with a horizon that is short whist not taking the aspects of stability and performance [9]. Figure 3 shows a general example of how model predication occurs and the general outcome that is expected and most likely achieved.

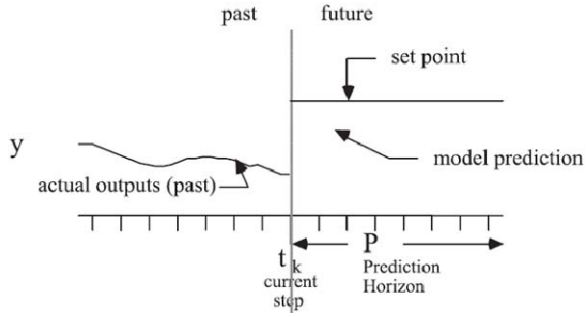


Figure 3: Nonlinear Model Prediction Control general prediction and outcome

The general outcome that one goes for is that for the open loop optimization. It is enough that at every interval, a decrease in the value function to give absolute stability and that can be implied on feasibility from the stability is achieved [10].

The major drawback is that the decreasing solution at every interval with optimization strategies interval are very expensive. Hence, if feasible solutions are delivered at every iteration and cost function decreases using the optimization strategy; then if no more intervals are available, the optimization can be stopped and at this time, the stability will still be guaranteed.

For both practically and theoretically for a good control solution that has been proven for model predictive control. When non-linear models are involved, this causes a much harder issue to tackle due to the control factors and computational difficulties [11].

There are a lot of challenges occurring when using Nonlinear Model Predictive Control and still a few unresolved questions such as: (1) Results of the Nonlinear Model Predictive Control output feedback exists but not always help in the real world problems or processes; (2) State estimations suitable strategies still needs to incorporated in the Nonlinear Model Predictive Control for a solution that is better for formulations; (3) Nonlinear Model Predictive Control applications in the industry is definitely rapidly growing, but none of the Nonlinear Model Predictive Control solutions include the critical stability constraints needed in nominal stability for the control theory [12]. Instead, the method used is to set an effective approximate of an infinite horizon by setting the horizon long enough [3]. For the NMPC theory, the differences with the industrial developments are still too large and needs the gap to be closed.

III. THE PROPOSED METHOD

Discrete Element Method (DEM) has been widely accepted not only in comminution but also to simulate environments where granular materials are used. In the present day, DEM technique is very strong and versatile tool for the development of mechanistic modeling in comminution [13].

Characteristics of DEM used in this research include: (1) simulates particle motion through Newton's equation of motion; (2) simulates particle collisions or contact through the laws of contact;

(3) calculations for the efficiency in relation to energy; (4) equipment design and optimization; (5) size reduction simulations in machines with comminution [14].

Revealing of the nature of the collisions energy spectrums is one of the crucial successes of the (DEM) Discrete Element Method. The concept of incremental breakage was also developed as DEM simulations provided information on the high number of weak collisions inside tumbling mills. [15].

EDEM® allows simulations with complex geometries that can be imported from third party software and it also attributes motion to both geometry and particles that are created within the software.

TABLE I.

EDEM simulations - Parameters of Material Used		
Specific Gravity (kg/m³)	*Shear Modulus (GPa)*	*Possion's ratio*
7800	0.1	0.1

One of the greatest advantages of EDEM® is its powerful capacity of post processing data, which allows the extraction of almost every type of information used during the simulation stages [16].

Secondly, the contact parameters were changed to compensate the ore that exists in the simulations. From prior researches, it is a reasonable assumption as the reaction of the steel balls are affected by the presence of particles of ore in the charge. The new estimated contact parameters represent presence of particles in contact of ball to ball collisions, regardless of the fact that they do not exist in the simulations. The differences of parameters or values are quite clear when used to represent ball-ball and ball-wall contacts, both being steel-steel surfaces.

TABLE II.

Parameters of Ore Particle Existence and Steel-Steel Surfaces		
Contact Parameters	***Steel-Steel Surfaces***	***Estimated Values (Accounting for presence of Ore)***
Restitution Coefficient	0.7	0.65
Static friction Coefficient	0.2	0.24
Rolling friction Coefficient	0.01	0.48

However, as the amount of generated information or data becomes very large in some cases, the data resulting from DEM simulations must be properly treated. A computational tool - LTM Analyst analyzes the performance of ball mills and to showcase data including: (1) power drawn; (2) particles velocity; (3) collisions magnitude and frequency; (4) the function of time in relation to the particles' position [17].

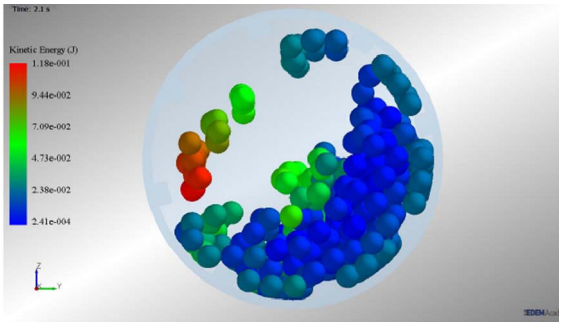


Figure 4: Snapshot of simulation which shows kinetic velocity over the grinding balls

The LTM Analyst uses the variation of the centre of mass of charge to calculate the power of the mill [18]. EDEM® will be used to record the information or data of all particle positions, including the rotation and speed and any tangential loss of energy.

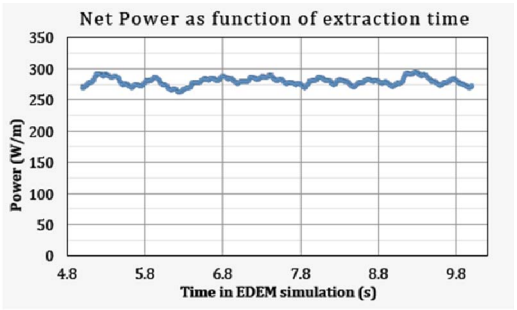


Figure 5: Power drawn of the ball mill

It will be easier to identify differences on breakage patterns, especially using larger diameter balls. This resulted in more intense breakage, especially of coarser and middle size particles. Figure 6 paints a clearer picture below. Left – d_b = 15mm; Right - d_b = 40mm

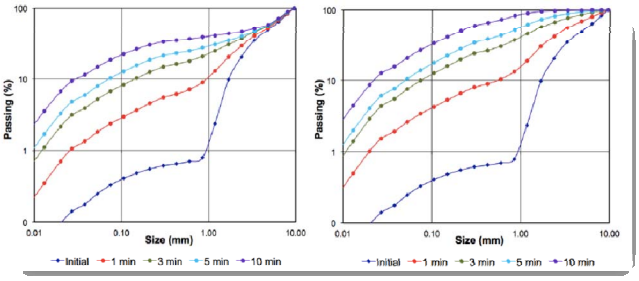


Figure 6: Size analyses of each step of grinding in a batch (J=30%, U=100% and φc=75%)

Smaller grinding media are not able to break coarse particles because they cannot provide the energy needed to break them during impact [19]. This can be confirmed by observing from the figure 7. The difference on breakage of the top size class (8mm) reached 66% when choosing 40 mm over 15 mm balls.

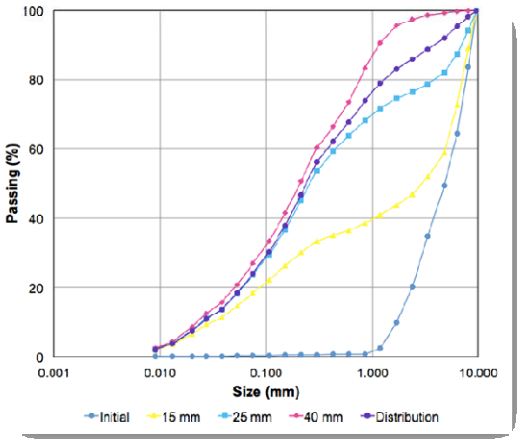


Figure 7: Size distribution of various ball sizes for 10 minutes of grinding (φc=75%, U=100% and J=30%)

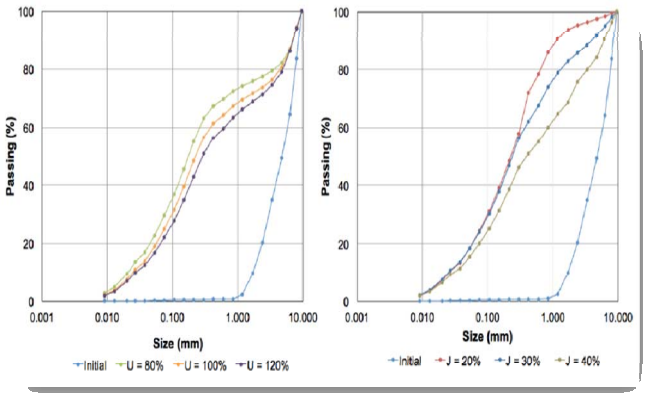


Figure 8: 25mm balls of grinding for 10 minutes - Powder filling effects on left and mill filling effects on right

The probability of impacts is much higher as there is a greater chance of balls making contact with the particles which have a smaller quantity.

In addition, the figure 8 (right) shows data regarding the effect of mill filling. Less mass (balls) resulted in more intense breakage, indicating that 20% of mill filling should be the optimal breakage condition.

IV. Experiment And Results

Ball variable is the variable that most constantly indicates a considerable effect on power. The ore charge became much finer as the power decreases. With regards to the ball size, the ball lifters capacity at elevating small balls, which is what dislocates the centre of gravity away from the mill centre. This also requires a considerable amount of energy which results in power draw as the grinding media diameter increases.

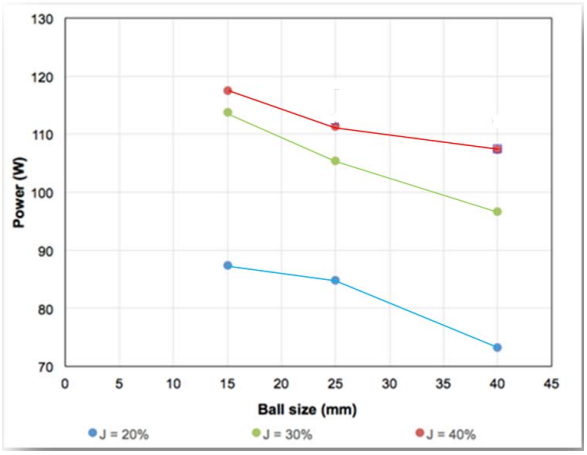


Figure 9: Power consumption with effects of mill filling

The effect of milling can be seen on figure 9, which indicates a relation between power consumption to the mass of the charge. There is a noticeable difference between 20% of mill filling and the other results independent of ball size.

The test using 40% of the mill filling demonstrated to consume much more energy when the grinding tests are compared.

Powder filling increases power consumption as it adds mass to the charge [20]. Figure 10 indicates there is a noticeable difference between tests with charge composed of balls plus ore powder from those compared with charge of steel balls only. The difference of power among tests with 80%-120% of powder filling almost cannot be distinguished, varying less than 5%.

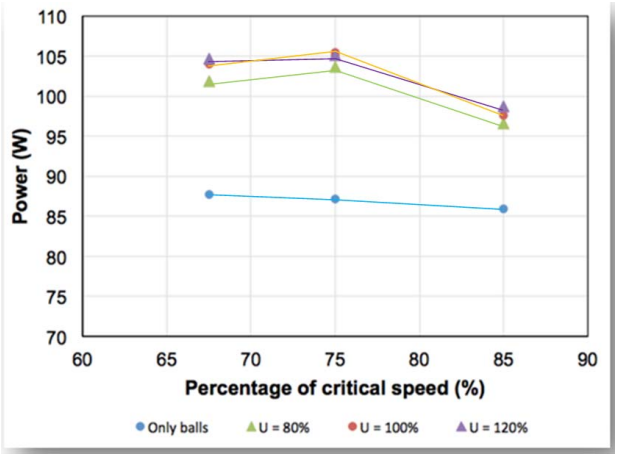


Figure 10: Power variation in respect to various powder fillings with critical speeds

From figure 10, it is quite evident that mill filling and the sizes of the ball have a major effect to power consumption. A stronger assistance of the ore towards power is achieved by having a low percentage of mill filling. The effects caused by these operation variables may change accordingly to the mill internal configurations.

(1) Operational variables have a huge effect on the power consumption factor in grinding and the particle breakage during the process. The process of grinding can be fully altered if one of variables happens to be changed or tampered with.

(2) Grinding media size is the variable that effects the consumption of power and particle breakage the most. The use of balls with a bigger diameter results in breakage with higher efficiency, this is much higher in particles that are coarse. This leads to consumption of power that is much less, however this effect may be linked to the high lifter used. Regarding the variables chosen, either quick breakage or milling efficiency can be prioritized. Bigger grinding media and slower mill speed provided the best grindability.

(3) Discrete Element Method solutions using the software, provided great insight into charge motion and showed to be sensitive to most changes in operational variables, but did have a drawback regarding the magnitude of power consumed.

V. Conclusion and Future Work

This paper addressed the performance of the grinding process in ball mills, with multiple variables concerned. The research has widened the options of the implementation of predictive control in grinding processes of ball mills. The economic performance enhancement can be increased by further studies and implementation of DEMs on a larger scale. DEM has shown to be a very powerful tool and with further research and studies on how to improve the tool to create real life processes in simulations without any drawbacks will be vital to the control of grinding processes in ball mills. Experimental work in the future with the use of lower rotational speed should be conducted to acquire the optimal operational conditions under which the mill will achieve performances of higher efficiency is of highest priority.

References

[1] C.W Steyn and P.G.R De Villiers, A Holistic Approach to Control and Optimization of an Industrial Run-of-Mine Ball Milling Circuit, Cape Town, IFAC Learning, 2014.

[2] Moritz Diehla,, H. Georg Bocka, Johannes P. Schlo, Rolf Findeisenb and Zolatan Nagyc, Real-time optimization and nonlinear model predictive control of processes governed by differential-algebraic equations, 2nd ed, 2005, [Online] Available: Knovel e-book.

[3] B.J.D Plessis. Grinding control strategy on the conventional milling circuit of Palaborwa Mining Company, Phalaborwa, Crossref Text, 2013, pp. 165-167.

[4] S. Morrell, Power draw of wet tumbling mills and its relationship to charge dynamics – Part 1: a continuum approach to mathematical modeling of mill power draw. Transactions of the Institute of Mining and Metallurgy, London, McGraw-Hill,1996**,** pp.43-53.

[5] B. S. Vladimir Monevand and Stefan Stoenchev. Grinding in Ball Mills: Modeling and Process Control. Minerals Engineering, vol. 12, 2012, pp. 101-105.

[6] P. Cleary. Predicting charge motion, power draw, segregation, wear and particle breakage in ball mills using discrete element methods. Minerals Engineering, 11th edition**,** 1998, pp**.** 1061–1080.

[7] G. Vallantyne, M. Powell and M. Tiang. Proportion of energy attributable to comminution. 11th ed, AusIMM Mill Operators' Conference, AusIMM, Hobart,Tasmania. 2012, pp. 25-30.

[8] R. M. Sarvalho, and L. M. Tavares. Predicting the effect of operating and design variables on breakage rates using the mechanistic ball mill model. Minerals Engineering, pp. 43-44, 2013.

[9] J. A Herbst and D. W Fuerstenau. Mathematical Simulation of Dry Ball Milling Using Specific Power Information. Trans. AIME, 1993, pp**.** 343-348.

[10] E. R. Cunha, R. M. Carvalho and L. M. Tavares. Mechanistic model of the vertical shaft impact crusher. In: 13th European Symposium on Comminution and Classification. Proceedings of the 13th European Symposium on Comminution and Classification, v.1**,** pp**.** 343 – 343, 2013.

[11] T. J. Napier, S. Morell and R. D. Morrisson. Mineral comminution circuits: Their operation and optimization, Indooroopilly, Qld, Julius Kruttschnitt Mineral Research Centre, The University of Queensland, June 1996.

[12] K. H. Erzadezaghh, A. Mirghasemi and A. A. Mohammadi. Numerical simulation of particle breakage of angular particles using combined DEM and Technology, Dubai, Science Direct, 2011, pp**.** 15-29.

[13] L.M Tavares and R.M Carvalho. Modelling ore degradation during handling using continuum damage mechanics. International Journal of Mineral Processing, 101 (1-4)**,** pp**.** 21-27, 2011.

[14] B.A. Wills, Mineral processing technology: An introduction to the practical aspects of ore treatment and mineral recovery, Boston, Butterworth-Heinemann, April 1997.

[15] D. Niebuhr, D. Pulm and W. Witt, Analysis, Conversion and Visualization of Particle Size Data combined with a Data Based Control System for Particulate Processes, Frankfurt, Oxford (OED) Learning, 1998.

[16] R. Morrison, B. Loveday, N. Diordievic, P. W. Cleary and P. Owen,. Linking discrete element modelling to breakage in a pilot scale AG/SAG mill.Advances in Comminution**,** London, Crossref Text, pp**.** 495–512, 2006.

[17] A. B. Makokha, M. H. Moys, M. M. Bwalya and K. Kimera. A new approach to optimising the life and performance of worn liners in ball mills: experimental study and DEM simulation. International Journal of Mineral Processing, vol. 84**,** pp. 221–227, 2007.

[18] R.P. King. Modeling and simulation of mineral processing systems, Oxford University,Butterworth-Heinemann, August 2001.

[19] Jeffery Williams,. Web-Based Economic Optimization Tools for Reducing Operating Costs Emerson Process Management [Online].Available from: www.emersonprocess-powerwater.com.], [Accessed 8th March, 2018].

[20] V. Adetola. Integration of real-time optimization and model predictive control. Milan, IFAC Learning, 2008.

An Ant Colony Hybrid Routing Protocol for VANET

Elias Khoza
Computer Systems Engineering
Tshwane University of Technology
Pretoria, South Africa
eliaskhoza@gmail.com

Chunling Tu
Computer Systems Engineering
Tshwane University of Technology
Pretoria, South Africa
tclchunling@gmail.com

Pius Adewale Owolawi
Computer Systems Engineering
Tshwane University of Technology
Pretoria, South Africa
OwolawiPA@tut.ac.za

***Abstract*— Vehicular Ad-hoc Network (VANET) is a definite form of Mobile ad-hoc Network (MANET), which delivers data communication in Vehicular environment using wireless transmission. Its key goal is to increase the service quality of intelligent transportation systems (ITS), such as road safety, logistics, and environmental kindliness as well as information exchange. Nowadays transportation systems are facing serious issues in terms of performance and efficiency of VANET applications, nevertheless these depend typically on the method in which messages are conveyed between the nodes. Finding a better routing protocol for dynamic VANET systems is one of the main challenges. This paper presents an Ant Colony Hybrid Routing Protocol (ACOHRP) to improve the service quality of ITS, by increasing the efficiency and reliability of vehicle traffic information message transmission. ACOHRP delivers high efficiency through better beginning of packet delivery ratio to end-to-end delay. A comparative study on the proposed ACOHRP and existing Dynamic Source Routing (DSR) protocol is conducted in a realistic scenario with VANET architecture, to demonstrate the performance of the proposed method. Simulation based testing is performed using Matlab with ACOHRP performing better than DSR in a dynamic environment of VANETs.**

Keywords—Ant Colony Optimization, Vehicular Ad-hoc Networks, Hybrid Routing Protocol.

I. INTRODUCTION

Currently VANET is gaining a lot of popularity among the industry as well as among the academic research community. This thought has been measured as the most valuable for improving performance and efficiency for future transportation. The operation of wireless communication technologies in vehicles as VANET has been growing quickly to improve road safety in intelligent transport systems (ITS). This kind of network permits the vehicles-infrastructure and vehicle-vehicle interactions in order for communication to take place. VANET is a type of MANET, whereby vehicles are denoted by nodes that are characterized through set of particular properties that changes speedily and continuously leading to an extremely dynamic topology. The routing protocol must support the high mobility, however, most MANET routing protocols are incompatible with VANET system. When deploying a routing protocol, certain properties need to be taken into consideration such as speed, security, geographic position, scope and the quantity of information supplied in order to enhance the movement of a signal. The inter vehicle network incorporates diverse wireless technologies such as 3G and 4G cellular, long term evolution (LTE) and LTE Advance.

This paper classifies existing routing protocols based on the architecture of VANETs [1], discusses the limitations, strength and strategies of each category. Based on qualitative comparison of performance and environmental feasibility, it is shown that geo-cast based routing and position-based routing are more accurate as compared to other routing protocols [2]. A new VANET routing protocol is then presented using Ant colony to structure the fast and reliable road traffic data transmission.

The remainder of this document is structured as follows, Section II briefly describes the related work of various VANET routing protocols with different strategies. In Section III, the proposed approach is presented in detail. Section IV presents the results analysis for three different scenarios. Section V concludes the paper.

II. RELATED WORK

A. VANET routing Protocols

VANET is an encouraging growing technology for subsequent generation of vehicles. VANET is offering a variety of applications, however the main concern is to discover an efficient routing protocol that is feasible to the highly dynamic VANET. For this challenge, several routing protocols were proposed which can be categorized into eight types.

1) Global Topology Routing Protocols

This kind of routing protocol needs the topology of all vehicles so that the information about links can be used to make routing decisions in the VANET. These protocols have advantages of discovering and maintaining the routes before the data transmission begins [2]. These protocols are categorized as follows: proactive, reactive, and hybrid protocols.

Proactive protocols are protocols that stores routing information of all the nodes within the entire network and it is table driven. The information is updated constantly because the nodes keep on changing its position every time [3]. The main

advantage of proactive protocol is the availability of information about the nodes in the routing table because there is no route discovery [4].

Reactive protocols are known as on-demand routing since there is no updated information about the topology, and routing information is retained only when required. This protocol is based on the principle of flooding the network with the route request and route reply messages [4]. It requires a route discovery mechanism unlike proactive because there is no information about the route maintained. It has an advantage of suitability of higher mobility scenario and for large scale networks [5, 6].

Hybrid protocol is the combination of both proactive and reactive routing protocols based on their features. Its main function is to reduce the delay in proactive protocols, and decrease the control overhead in reactive routing protocols [7].

2) Topology Free Routing Protocols

These protocols are based on position information for the moving nodes and they are also known as a geographic routing protocol. It uses GPS (Global Positioning System) to locate the nodes that are participating within the network. One of the advantages of these protocols is that the route discovery and route maintained are not required, unlike topology based routing [8].

3) Cluster Based Routing Protocols

These is the type of protocols are based on the principle of clustering in which group formation and cluster head selection are determining the process. One of the important advantages of this protocol is lowering costs and the delays in data packet delivery [9]. Clustering is also suitable in big scale distribution networks for easier management. However, each and every group is assigned a cluster lead, which is accountable for nodes, management of similar groups as well as amongst the other groups [10, 11]

4) Geo-Cast Based Routing Protocols

These protocols utilize GPS to learn about the position of nodes and it is a position based multicast routing. One of its benefits is to ensure that the packets are delivered within an identified geographical region from source node to all other nodes in the network. It is also considered as a multicast service within the identified geographic area. The operation of this type of routing uses multicast group within a specified geographic area to broadcast a packet by directing flooding technique. [12].

5) Multicast Based Routing Protocols

These protocols focus on transmitting packets within specific regions from single source to numerous targets [13]. Multicast routing adapt to networks properties by considering high mobility, high speed movement and frequent topology variations. Multicast protocols are beneficial to VANETs since their wireless nature permits a packet to be broadcasted all over the network within the same range [14].

6) Broadcast Based Routing Protocols

This is a protocol with numerous benefits in VANET such as distribution traffic, emergency, and weather, road situation amongst vehicles, and supplying advertisements, messages and unicast for a well-organized route [14]. The standard of operation is by flooding and independent node re-broadcast to others which guarantee the message delivery. The advantage of this type of protocol is to guarantee message delivery, easy to set up but with some limited nodes.

7) Delay Tolerant Routing Protocols

In the process of avoiding congestion and complexity, this type of routing is introduced in VANET with several partitions resulted in the high flexibility. The carry and forward method is utilized by vehicles in order to store data packets to the network. [15]. One of the characteristics of Delay Tolerant routing is Intersection-based Geographical Routing Protocol (IGRP) which has been recognized by most investigators for routing packets in city surroundings [16].

8) UAV Assisted Routing Protocols

This is an improved type of routing protocol completing the connected sections as well as advancing routing in order to have a world-wide vision for UAVs. [17]. One of the important benefits of this protocol is to overcome the presence of obstacle along the path with two categories namely: UVAR-G and UVAR-S. It also guarantees the packet delivery by sending data packets on the ground over connected vehicle. [18, 19].

B. The Structure and Behaviour of VANET

VANET and MANET have almost the same structures, but differ in high mobility of nodes, that makes regular technological variations. Thus, a vehicle can rapidly join or leave a group of vehicles in a little space of time, resulting in having little connectivity. Moreover, VANET supply broadband connectivity and technical resolutions with great accuracy [19]. When any vehicles enter the cluster zone, its default status will be a Cluster Member (CM) and the HELLO message will be exchanged with cluster header (CH) as shown in Fig. 1.

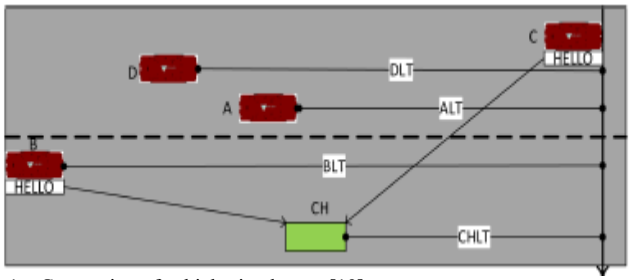


Fig. 1. Connection of vehicles in clusters [19]

1) Cluster creation

Vehicles change topology regularly and restricted lifetime of connections between vehicles because of high mobility. This will allow the vehicles to move based on predefined methods by road infrastructure as well as traffic laws [20]. In order for better communication of application of specific data, a cluster leader is nominated to construct and keep the structure of the clustering mechanisms. Cluster creation permits the division of networks into smaller segments for easy management [21].

Vehicles that are in the same cluster connection can be able to exchange information efficiently for a period of time on normal speed. [21]. The nodes cluster will be built based on two standards identified in Fig. 1 and Fig. 2. If the cluster is not yet recognized, and there are at least two vehicles of which their speed and path are checked from Fig. 2 (a) [22]. The selection of clustered (CH) will depend on the path of the route segment amongst vehicles that should be suitably long enough to form a connection and exchange information [22]. On the other standards if the cluster is already created, the request will be broadcasted for integration with path and speed consequently as shown in Fig. 2 (b).

2) Cluster change for a simple node

The migration of a cluster for one vehicle to another is carried out by the route and speed changes. The vehicle changes will suit any new vehicle added to the cluster. If the vehicles approaches to the end of mutual segment, then it needs to search for a new cluster using cluster-head [22]. If the vehicle changes suddenly in terms of path or speed, then it will search for a new cluster by returning a warning message to the old cluster. [23]. The vehicle must choose its own successor before leaving the cluster if needs to change speed or path and as a result it must relate the pathways of all vehicles. [23, 24].

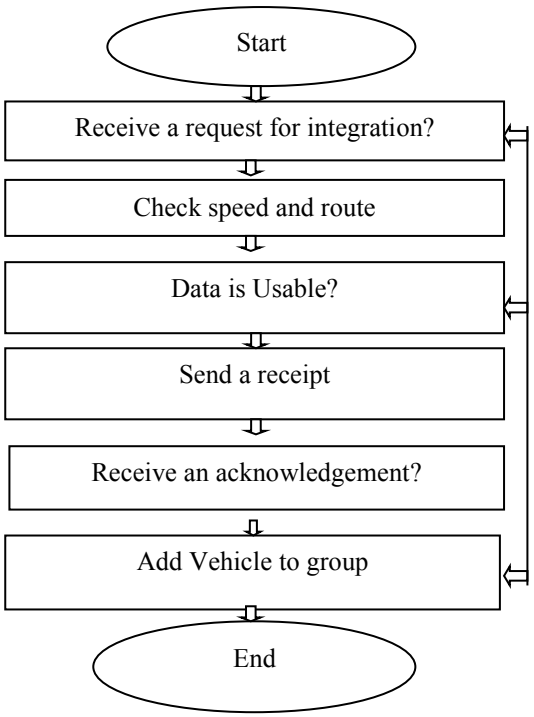


Fig. 2. (a) Establish a connection for Vehicle [3]

III. ACO HYBRID ROUTING PROTOCOL (ACOHRP) FOR VANET

This routing technique enable the vehicle to communicate different information with other vehicles efficiently for a certain time depending on the average speed and route of the vehicles that are in connection. Fig. 2(a), represents a group of vehicles that will be checked against route and speed in the common route segment. On the other hand, Fig. 2(b), elaborates the broadcast of vehicles against path and speed respectively. Clusterhead creation is responsible for sending an acceptance signal and awaits an acknowledgement [24].

1) Ant Colony Technique

Ant Colony (ACO) can be described as metaheuristic protocol that is influenced by scavenging conduct of ants. It use the Pheromone (hormone) which is deposited and identified by ants when they passing along the paths. [24]. Pheromone captivated ants, which cause more ants to get attracted to the same path. The ACO can point out a selection technique which raises an issue by iteratively trying to improve a candidate solution with respect to a particular measure of quality. In VANETs, ants can be denoted as special packets and rules that can be configured based on the algorithm for the packets [25, 26].

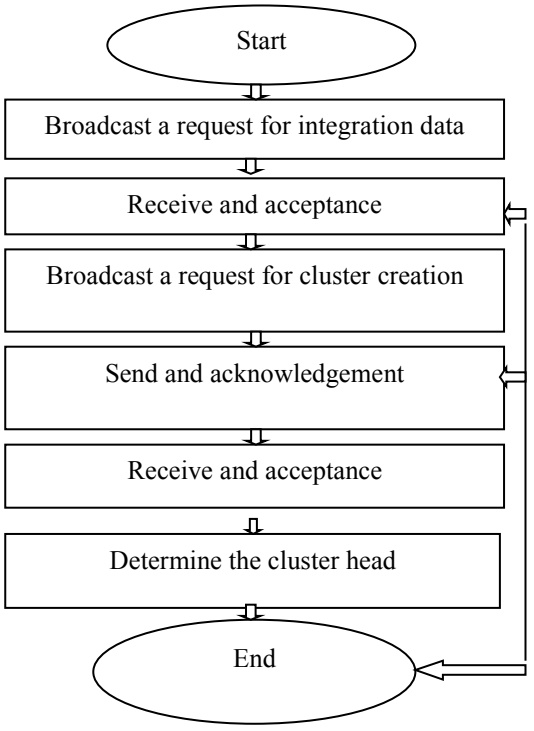


Fig. 2. (b) Steps to add new vehicle in a cluster-head [3]

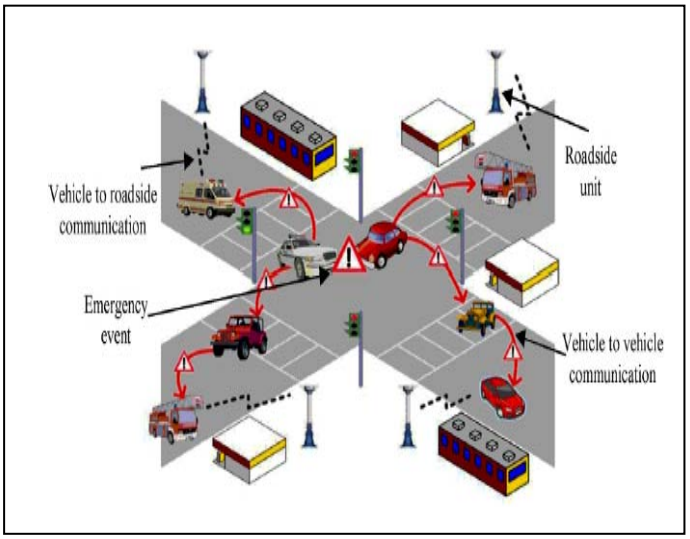


Fig. 3. VANET Scenario [4]

2) ACO structure for VANET

As earlier mentioned, the method is based on traffic information in order to enable the communication between the vehicles, and to achieve this, the Ant Colony is employed to signify knowledge and vehicles traffic information. This section describes the architecture of the proposed system as shown in Fig. 3. ACO consists of numerous processes such as configure nodes, network initialization, source and destination allocation, data transmission and performance analysis [25, 26].

Configure nodes – the nodes are organized as mobile vehicular nodes as specified from VANET.

Network initialization – in this scenario, the nodes are well-defined and it is done with specified area.

The data transmission uses DSR path finding protocol that is located in Zone based ACO whereas Source and destination nodes are assigned by the users.

The method focuses on traffic information to simplify communication amongst vehicles, for this reason ant colony is proposed to represent vehicles information and knowledge.

3) Dynamic Source Routing (DSR)

DSR is a sensitive path-finding scheme which does not need periodic HELLO packets and warning signal. The DSR protocol technique is deluging the packets within the network by using route request. The node responses over destination and conveys route traverse in its Cluster Header. DSR is composed of two techniques which collaborates and works to permit route maintenance and route discovery [27]. Route maintenance is done by the propagation of the Route Error message (RRER). Route discovery is done whenever a source node needs to transfer any packet to an end point node, which starts by consulting its source cache [28]. DSR is designed especially for the usage in a multi hop ad-hoc network of mobile nodes. This permits a network to behave entirely self-organized and self-configuring without using any existing infrastructure. This protocol uses no periodic routing messages, therefore it avoids large routing updates and decreases network bandwidth overhead [27, 28].

4) Improved Ant Dynamic Source Routing

ANT-DSR is a reactive technique that uses proactive route protocol over a constant validation of its stored routes. In this scenario when packets are transmitted, proactive methods are used within the network and reactive methods between networks [29]. The performance metric is improved when ant-net algorithm is applied to DSR protocol. This technique increases life span of a listed route in the VANET network [29].

IV. Results and Discussion

In this subsection, the performance parameters as well as efficiency of network estimation are discussed. These parameters are:

a. *Routing Overhead* – for carrying routing information, the same bandwidth used by data packets is added to the user transferring data with the nodes exchanging routing information. The overhead parameter estimates additional traffic in which routing protocol produces. In order to accomplish routing tables. It is usual to have higher traffic loads than other algorithms, since all nodes in the network re-broadcast messages without processing through a calculation.

b. *Throughput* - Fig 4. Indicates the throughput values of ACOHRP and DSR. The outcome confirm the conclusion of the traffic information results. Throughput increase gradually for both protocols with an increasing number of vehicles. It is seen that ACOHRP outperforms DSR with the highest throughput.

c. *End to End delay* – Fig 5. Represents the average end to end delay of ACOHRP and DSR. From the observation, it can be seen that the average end to end delay gradually improves for both protocols as zone radius increases, due to period consumed by route discovery mechanism. Also with the increase of the zone radius, the relay of information packets consumed a lot of time and will result in the increment of delay packets relay by hops. It is observed that ACOHRP has the highest delay.

By making use of Matlab program, the results below were obtained using simulations. The next measurements have been executed over simulations for the routing protocols that

have been concisely described. Routing overhead, end-to-end delay and throughput. The values in the figures are obtained by simulation average set for each vehicles.

Table I: SIMULATION PARAMETERS

Parameters	**Value / Range**
Mobility Model	MATLAB
Network Traffic Connections	3 CBR Connections
Simulation Zone Area	20m X 20m
Number of Nodes	100, 200
Fading	Nil
Velocity of each Node	Constant
Speed	Up to 20 m/s
Transmission Range	< 100 m
Routing Protocols	DSR
Data Packet Size	512 bytes

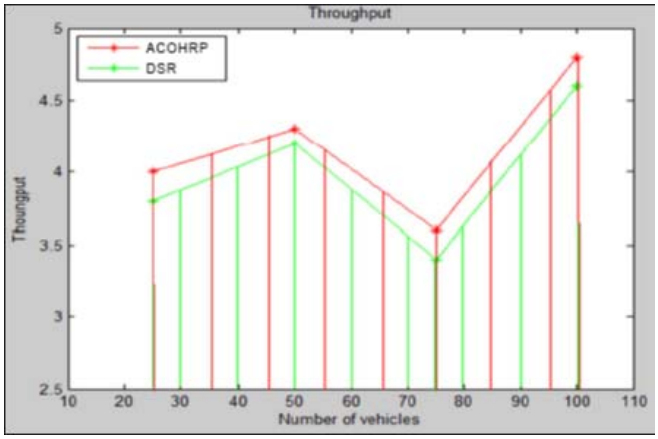


Fig.4. Comparison of throughput VS Vehicles

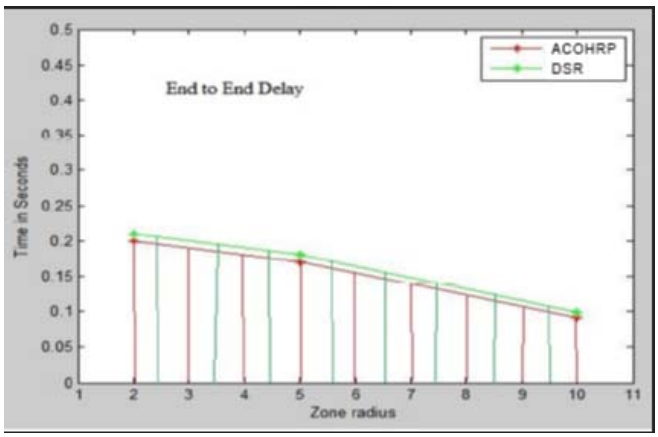


Fig.5. Comparison of End to End delay VS Zone radius

V. CONCLUSIONS

The networks of VANET are featured by high mobility vehicle, in which vehicle can connect and disconnect from the cluster network in a very limited period, and frequent topology changes occurred. Low latency multipath routing structure has been introduced by using an ACO method for vehicular network communication. The different routing protocols are required in order to permit the proper routing of packets to their final destination.

The paper, is focusing on traffic information as well as ACOHRP in order to allow suitable routing of packets from source to their final destinations. The ant colony is used in vehicles to enable the analysis of information acquired from traffic. Furthermore, the ACOHRP has excellent flexibility for routing in different ad hoc networks. It also includes properties such as dynamic topology, efficient path selection as well as evaluation of link transmission quality. Three factors are measured to calculate discovered paths, namely, routing overhead, end-to-end delay and throughput. This approach optimize routing by improving road service performance and can reduce significantly the delivery time. The results show that ACOHRP performs better as compared to DSR routing approaches. The new proposed ACOHRP will be a better solution to deal with all kinds of traffic scenarios and can outperform well when evaluated using VANET metrics in dynamic environment. For future scope, hybrid protocol can be developed in order to overcome all drawbacks of existing protocols.

ACKNOWLEDGMENT

This work is based on the research supported by the National Research Foundation of South Africa (Grant Number 99383).

REFERENCES

[1] C. R. Nair, " Analysis and comparative study of topology and position based routing protocols in VANET," International Journal of Engineering Research and General Science, Vol 4(1), pp. 43–52, January – February, 2016.

[2] S. Saha, U. Roy, and D. D. Sinha, " Performance comparison of various Ad-Hoc routing protocols of VANET in Indian city scenario, " American international journal of research in science, technology, engineering and mathematics, Vol 5(1), pp. 49–54, 2014, February.

[3] H. Toulni, and B. Nsiri, " A hybrid routing protocol for VANET using Ontology," The international conference on advanced wireless, information, and communication technologies (AWICT 2015), Vol 73, pp. 94–101, 2015.

[4] B. T. Sharef, R. A. Alsaqour, and M. Ismail, " Vehicular communication Ad-hoc routing protocols: A survey, " Journal of Network and computer applications, Vol 40, pp. 363–396, 2014.

[5] A. Abuashour and M. Kadoch, " Performance improvement of cluster-based routing protocol in VANET, " Vol 5, pp. 15354–15371, 2017.

[6] S. S. Joshi, and S. R. Birada, " Communication framework for jointly addressing issues of routing overhaed and energy drainage in MANET," Twelfth international multi-conference on information processing, Vol 89, 57–63, 2016.

[7] B. Sahadev, K. S. Zade, and S. H. Sheikh, " Survey on realistic simulation for comparison of network routing protocol in VANET," 2nd National Conference on Innovative Paradigms in Engineering & Technology, pp. 26-29, 2013.

[8] M. Ren, L. Khoukhi, H. Labiod, J. Zhang, and V. Veque," A mobility-based scheme for dynamic clustering in vehicular ad-hoc networks (VANETs), " Vehicular Communications, Vol 9, pp. 233–241, 2017.

[9] W. Angeles, V. P. Borin, A. Munaretto, and M. Fonseca, " The impact of propagation models in the performance of ad-hoc routing protocols for Urban VANET, " 2016 IEEE 84th Vehicular technology conference, 1–5, 2016.

[10] P. Mutalik, S. Nagaraj, J. Vedavyas, R.V. Biradar, and V. G. C. Patil, " A comparative study on AODV, DSR and DSDV routing protocols for intelligent transportation systems in metro cities for road traffic safety using VANET route traffic analysis (VRTA), " 2016 IEEE international conference on advances in electronics, communication and computer technology, 383–386, 2016.

[11] N. Goel, I. Dyani, and G. Sharma, " A study of position based vanet routing protocols," International conference on computing, communication and automation, pp. 655–660, 2016.

[12] I. Basaran, and H. Bulut, " Performance Comparison of non Delay Tolerant Vanet Routing Protocols," International workshop on urban mobility & intelligent transportation systems, pp. 1–6, 2016.

[13] P. S. Manjunath, and N. Reddy, " Assessment of VANET routing protocols in safety paradigm", pp. 183–188, 2015.

[14] B. Hamid, and E. E. Mokhtar, " Performance analysis of the vehicular ad-hoc networks (VANET) routing protocols, AODV, DSDV and OLSR," pp. 1 – 6, 2017.

[15] A. Datta, " Modified Ant-AODV VANET routing protocol for Vehicular Ad-hoc Network, " IEEE, pp. 1 – 6, 2017.

[16] A. Abuashour, and M. Kadock, " Performance improvement of cluster based routing protocol in VANET, " IEEE Access, Vol 5, pp. 15354 – 15371, 2017.

[17] W. Farook, M. Ali-Khan, and S. Rehman, " A multicasat routing protocol for automnomous military vehicles communication in VANET, " Proceedings of 2017 14th international Burban conference on applied science and technology, pp. 699 – 706, 2017.

[18] S. Benkirane, S. Mostafa, M. L. Hasnaoui, and A. Beni-Hssane," A new comparative study of Ad-hoc routing protocol AODVand DSR in VANET environment using simulation tools," 2015 15th International conference on intelligent system designs and applications , pp. 458 – 461, 2015.

[19] G. Yan, and D. B. Rawat, " Vehicles-to-vehicle connectivity analysis for vehicular ad-hoc networks, " Ad-hoc Networks, Vol 58, pp. 25 – 35, 2017.

[20] S. Khakpour, R. W. Pazzi, and K. EI-Khatib, " Using clustering for target tracking in vehicular ad-hoc networks, " Vehicular Communications, Vol 9, pp. 83 – 96, 2017.

[21] J. Skiles, and I. Mahgoub, " A geographical hybrid solution for inter-vehicular communication in VANET, " IEEE, pp. 250 – 255, 2016.

[22] J. Sathiamoorthy, B. Ramakrishnan, and M. Usha, " Design of proficient hybrid protocol for efficient route discovery and secure data transmission in CEAACK MANETs, Journal of information and security and applications, Vol 36, pp. 43 – 58.

[23] J. Shen, C. Wang, A. Wang, X. Sun, S. Moh, and P. Hung, "Organized topology based routing protocol in incompletely predictable ad-hoc networks, " Computer Communications, Vol 99, 107 – 118, 2017.

[24] O. S. Oubbati, A. Lakas, F. Zhou, M. Gunes, N. Lagraa, and M.B. Yagoubi, "Intelligent UAV-assisted rounting protocol for urban VANETs, " Computer Communications, Vol 107, 93 – 111, 2017.

[25] R. Skinderowicz, " An improved ant colony system for the sequencial ordering problem, " Computers and Operations Research, Vol 86, pp. 1 – 17, 2017.

[26] Y. Meraihi, D. Acheli, and A. Ramdane-cherif, "QoS performance evaluationof AODV and DSR Routing Protocols in City VANET Scenarios, "The 5th International Conference on Electrical Engineering – Boumerdes, 1 -6, 2017.

[27] S. Chatterjee, and S. Das, "Ant colony optimization based enhanced dynamic source routing algoithm for mobile Ad-hoc network, "Information sciences, pp. 67 – 90, 2015.

[28] R. Skinderowicz, "An improved Ant Colony system for the sequencial ordering problem", Computers and Operations research, pp. 1 -17, 2017.

[29] A. Louati, S. Elkosantini, S. Darmoul and L. Ben Said, " A Case Based Reasoning System to Control Traffic at Signalized Intersections" Procedia Computer Science , Vol 5, pp. 149 – 154, 2016.

Improving Association Rule Mining Using Clustering-based Discretization of Numerical Data

Swee Chuan Tan
School of Business
Singapore University of Social Science
463 Clementi Road, Singapore
jamestansc@suss.edu.sg

***Abstract*—Association rule mining is an important data mining technique that help discover interesting attribute relationships that are useful for decision making. Most association rule mining methods use item-set manipulation approach, whereby data type must be categorical in nature. When a dataset contains numerical attributes, they will need to be discretized before rule mining. At the moment, most unsupervised data discretization methods do not account for data distributions, and users have to try different methods and discretization settings in order to improve rule mining results. In this paper, we propose using TwoStep clustering for data discretization. Unlike simple discretization methods, TwoStep automatically determines the discretization intervals by taking into account the unique data distribution property of each attribute. In our experiments, we evaluated the performance of Apriori algorithm based on four datasets, whereby each dataset was pre-processed using TwoStep and three other commonly used discretization methods. Our results show that TwoStep produced the greatest number of high-quality rules, as compared to common discretization methods.**

Keywords— association rule mining, numerical attributes, discretization

I. INTRODUCTION

Association Rule Mining (ARM) is a data mining method [1] that discovers interesting and useful relationships hidden in data [3, 5, 6]. One common application of association rule mining is market basket analysis [14] in a supermarket setting, where products frequently purchased together are identified, and the interesting information gained [12] can be used to improve pricing and store layout, or even for formulating promotional strategies such as cross-selling and product bundling. In some cases, the algorithm can be extended to discover repeated sequence of events, where sequence rules [2] are used to predict what the customer is going to buy next, and when the next purchase will be made [15].

When applying association rule mining in practice, most ARM methods has a number of limitations [9, 10]. The most common limitation is that they require all attributes to be categorical in nature. Therefore, datasets with numerical attributes will first have to be discretized [7] before rule mining. Unfortunately, common discretization methods suffer from two limitations. Firstly, user must determine the correct number of intervals (a.k.a., bins) to use. Using too few intervals will under-represent the data and result in loss of information. On the other hand, using too many intervals will result in suboptimal representation of data, resulting in ineffective binning. Secondly, common discretization methods make explicit assumptions about data distributions; and they do not work well when their assumptions are violated.

As an example, a straight-forward discretization method is the Equal-Width Binning (EWB) method [8]. This method divides the range of a data attribute A into k equal intervals. However, EWB has two limitations. Firstly, the user must determine an appropriate value of k so that it can effectively and efficiently capture the data distribution. Secondly, EWB tacitly assumes that A has a continuous uniform distribution. When A is not uniformly distributed (e.g., with outliers or skewness), certain EWB intervals may capture the majority of data points while other intervals may capture the remaining few outliers.

The purpose of this paper is to address the above two limitations of common discretization methods by proposing the use of TwoStep Clustering [13]. This approach is promising for two reasons. Firstly, TwoStep Clustering (TSC) uses Bayesian Information Criterion (BIC) [11] to automatically determine the number of clusters, which partitions a continuous attribute into a set of non-overlapping intervals. The number of clusters in this case is the number of bins to be used for discretization. This does away with the need for a user to specific the value of k, which cannot be determined *a priori*.

Secondly, unlike traditional method such as EWB, the discretization bins generated by TSC do not have uniform width and frequency. The bins are generated to represent each specific data distribution so as to mitigate loss of information. Hence, there is less concern about how an attribute data distribution affects the discretization process.

The rest of this paper is organized as follows. Section II reviews common data discretization methods. Section III describes the proposed TSC approach for data discretization. Section IV presents the experimental setup and compares the performance of the proposed method versus that of common methods. Section V concludes this paper.

II. RELATED WORK

In this section, we will first review three common data discretization methods that have been used for association rule mining. Then, we examine some key issues in these methods.

A. Common Discretization Methods

Here we look at three discretization methods [8], namely (i) Equal Width Binning, (ii) Equal Frequency Binning, and (iii) Mean and Standard Deviation Binning.

Given a numerical attribute A containing n sorted values a_j, such that $a_1 \leq a_2 \leq \ldots \leq a_n$, where $j \in Z$: $j \in [1, n]$. The Equal Width Binning (EWB) method divides A into k bins, namely $B_1, B_2, \ldots, B_k$, where k is a user-specified parameter. Each bin has a fixed width $w = (a_n - a_1)/k$, and the interval

boundaries of each bin B_i are $[a_1 + (i - 1) \cdot w, a_1 + i \cdot w)$, where $i \in Z$: $i \in [1, k]$.

For Equal Frequency Binning (EFB) method, the values of A are divided into k bins of fixed size, where each bin contains n/k instances. Assuming that n is divisible by k (i.e., $n|k$), then any bin B_i contains a set of instances ordered from instance $a_{(i-1) \cdot n/k+1}$ to instance $a_{i \cdot n/k}$.

For Mean and Standard Deviation Binning (MSDB), the mean (m) and standard deviation (sd) of the values in A are computed. Then the data values are divided into three bins, where the first bin contains instances with values less than $m - sd$; the second bin contains instances with values within $m \pm sd$; and the third bin contains instances with values greater than $m + sd$.

B. Issues of Common Discretization Methods

When using the EWB and EFB methods, one common issue is the need to decide on the value of k, which is the number of bins to use. The general rule of thumb is to use the least number of bins possible to capture the largest possible amount of information from the data. If too many bins are used, the data distribution is not efficiently represented. On the other hand, too few bins cannot represent the data adequately and important information of the data will be lost after discretization.

In practice, it is not easy to decide on the number of bins without actually looking at the data distribution specifically. Usually, a histogram can be used for this purpose and the user can then decide upon the number of bins to use in order to best represent the data distribution. Unfortunately, different users may interpret the same histogram differently and therefore use different number of bins without any objective basis.

Finally, the MSDB approach assumes that the data is normally distributed. Any violation of this assumption, can result in MSDB not producing bins that capture the data distribution accurately. For example, a highly skewed distribution or the presence of outliers can result in a distortion of central tendency, resulting in inaccurate representation of the data distribution.

In the following, we will use two examples to illustrate the strengths and weaknesses of these three methods.

Figure 1 shows a continuous attribute that contains three equally-sized underlying normal distributions, namely D1, D2 and D3. It can be seen that D2 has the widest spread while D3 has the narrowest spread, and D1's spread is between that of D2 and D3. Since each of D1, D2 and D3 has the same number of instances, EFB is able to discretize these three distributions into three perfectly discretized intervals. Figure 1 also shows that EWB is particularly sensitive to attributes that contain distributions with uneven spreads. It can be seen that EWB assigns wider-than-required intervals for D1 and D3, which have narrower spreads; and EWB assigns a narrower-than-required interval for D2, which has a much wider spread.

Finally, we can see that MSDB does not work well here due to the violation of normal distribution assumptions, which assume a single bell curve and not a multimodal distribution such as the one shown in Figure 1.

Figure 2 shows another continuous attribute that contains three underlying normal distributions that have the same spread but different number of instances in their respective distributions. These distributions are namely D4, D5 and D6. It can be seen that D5 has the most number of instances while D6 has the least number of instances, and the number of instances in D4 is between that of D5 and D6. Since D4, D5 and D6 have the same spread, EWB is able to discretize these three distribution into three perfectly spaced intervals. However, EFB is sensitive to attributes that contain distributions with uneven sizes. It can be seen that EFB assigns wider-than-desired intervals to account for D4 and D6, which have less instances; and EFB assigns a narrower-than-desired interval to account for D5, which has the most number of instances.

Finally, we can see again that MSDB does not work well with this multimodal distribution.

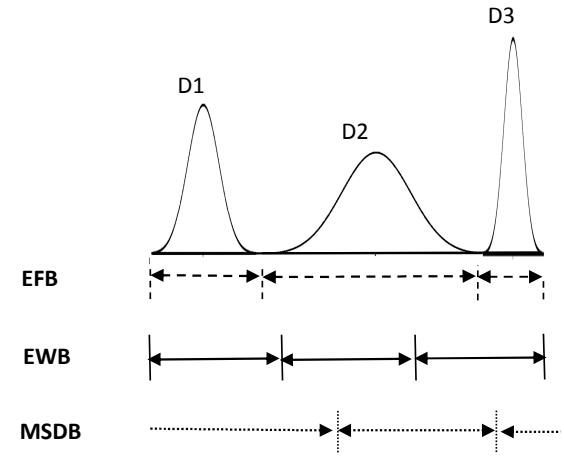


Fig. 1. By using the optimal number of bins (i.e., three bins) for three distributions D1, D2 and D3, all three binning methods have different results. In this case EFB has the best characterization of the data.

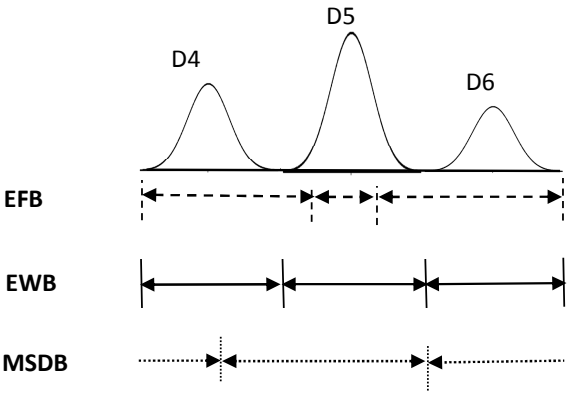


Fig. 2. By using the optimal number of bins (i.e., three bins) for three distributions D4, D5 and D6, all three binning methods have different results. In this case EWB has the best characterization of the data.

III. Proposed Method

To address the issues with common discretization method, we propose the Automatic Cluster-based Discretization (ACD) method. For each continuous attribute A of a dataset, we apply TSC to cluster the instances in A. Using Bayesian Information Criterion, the TSC method can

automatically determine the number of clusters to be used, and assign each data instance to a cluster.

In general, if A has a k-modal distribution, then TSC aims to partition the data instances in A into k disjoint clusters, where k is automatically determined using BIC. In other words, each data instance belongs to one and only one of the clusters. In the process of generating the clusters, each cluster is generated such that it contains as many similar values as possible; and the values between any two different clusters are to be as dissimilar as possible. In other words, the intra-cluster variance is to be minimized, and the inter-cluster variance is to be maximized.

Following this approach, a continuous variable A is partitioned into k non-overlapping bins: B_1, B_2, …, B_k. For any bin B_i, it represents a closed interval in which all instances with values that fall within this interval are grouped under bin B_i.

The advantages of using TSC are two-fold: (i) there is no need for the user to specify k as it is automatically determined by BIC; and (ii) TSC objectively creates the bins based on the data distribution property, and avoids relying on subjective human judgement to decide on the number of bins to use.

IV. Experimental Setup and Results

This section first presents the experimental setup, and then it evaluates the effectiveness of using the proposed method versus that of common discretization methods.

A. Experimental Setup

For our experiments, we use four real-world datasets, namely Wine, Abalone, Iris and Pima. They are taken from the University of California, Irvine (UCI) Machine Learning Repository website [4].

The Wine dataset contains results of 13 measurements used in a chemical analysis of three types of wines grown in Italy. The Abalone dataset contains abalone characteristics such as sex, measurements of physical dimensions, and various forms of weights. The Iris dataset contains sepal and petal measurements of three sub-species of Iris flowers. The Pima dataset consists of diabetic classification of Pima Indians in relation to attributes such as the number of previous pregnancies, body mass index, insulin level, age, etc. Additional details of these datasets are shown in Table 1.

TABLE I. Details Of Four Real-World Datasets Used In The Experiments.

Dataset	Data Size	Number of Attributes
Wine	178	13
Abalone	4177	8
Iris	150	6
Pima	768	8

Because these datasets contain largely continuous attributes, we preprocessed the continues attributes in each of these datasets using four methods, namely EWB, EFB, MSDB, and the proposed ACD. The software used is the IBM SPSS Modeler ® version 16. Throughout the experiments, default settings are used to ensure consistency and fairness when comparing and evaluating the results.

B. Evaluation

To evaluate the effectiveness of the data discretization methods, we first evaluate each discretization method by looking at the total number of Apriori [1] rules generated. We call this the Quantity Criterion. A good discretization method will produce a dataset such that the association rule mining algorithm is able to generate as many frequent itemsets as possible, thus resulting in a high number of association rules.

Based on the Quantity criterion, Table 2 shows the number of Apriori association rules generated from each of the datasets pre-processed by each of the four data discretization methods. Boldfaced entries indicate the highest number of rules generated, and underlined entries indicate the second highest number of rules.

TABLE II. Number of Apriori association rules generated from each of the datasets pre-processed by the four data discretization methods.

Dataset	EWB (Equal Width)	EFB (Equal Frequency)	MSDB (Standard Deviation)	ACD (TwoStep Clustering)
IRIS	11	<u>31</u>	30	**90**
WINE	3	227	<u>334</u>	**1973**
ABALONE	1	**2963**	1	<u>1817</u>
PIMA	26	<u>32</u>	28	**220**
Total	41	<u>3253</u>	393	**4100**

As can be seen, ACD is the discretization method that generates the highest number of Apriori rules in three out of four datasets. Even in the Abalone dataset, ACD generated the second highest number of rules.

It should be noted that the results in Table 2 are generated based on the same parameter settings of the Apriori association rule mining method. This suggests that ACD tends to create bins that result in more frequent itemsets being generated, and translates to the production of more association rules.

Table 2 also suggests that the default settings of EWB and MSDB do not work well for the Wine and Abalone datasets. This is evident from the very few (i.e., three or less) rules generated from these datasets. This problem may be alleviated by adjusting the settings of EWB and MSDB. On the other hand, there is no need to do so when using ACD.

To complement the Quantity Criterion, we then evaluate the results based on the Quality Criterion, where we examine association rule quality by evaluating the confidence and support of top-N rules being generated. In particular, we study the mean rule-confidence and mean rule-support percentage figures of up to 20 rules that attained the highest rule confidence.

Table 3 compares the association rule mining results based on four discretization methods applied across four datasets. Each entry in this table represents the mean rule support percentage, followed by the mean rule confidence

percentage of up to 20 possible rules that attained the highest rule confidence. Boldfaced entries represent the highest scores, while underlined entries represent the second highest scores.

In Table 3, it can be seen that ACD attains the highest rule support and rule confidence in most datasets. The only exception is the Iris dataset, where ACD attains the second highest mean rule support of 17.6%, which is only marginally below the highest score of 17.9% achieved by MSDB.

TABLE III. THE QUALITY OF APRIORI ASSOCIATION RULES GENERATED BASED ON UP TO 20 POSSIBLE RULES THAT ATTAINED THE HIGHEST RULE CONFIDENCE. EACH ENTRY IS REPRESENTED BY (S, C), WHERE S IS THE MEAN RULE SUPPORT, AND C IS THE MEAN RULE CONFIDENCE.

Dataset	EWB (Equal Width)	EFB (Equal Frequency)	MSDB (Standard Deviation)	ACD (TwoStep Clustering)
IRIS	16.8, 96.8	15.3, **100**	**17.9**, **100**	17.6, **100**
WINE	13.5, 90.2	12.9, **100**	12.6, **100**	**21.2**, **100**
ABALONE	8.8, 85.2	14.9, **100**	8.8, 85.2	**22.2**, **100**
PIMA	12.49, 88.5	13.1, 91.0	14.4, 88.0	**14.8**, **100**
Total	51.4, 360.6	56.2, 391.0	53.7, 373.1	**75.7**, **400**

V. CONCLUDING REMARKS

In this paper, we have shown that an automatic clustering-based discretization approach for data discretization is a promising way to improve association rule mining quality. In particular, we have used an existing clustering method known as the Two-Step Clustering, to help automatically determine the number of bins to be used. In additional, it is better able to accommodate to the data distribution property. Our experiment results show that ACD generally outperforms common binning methods by a fair margin and is therefore a good alternative method for data discretization before association rule mining.

REFERENCES

[1] R. Agrawal, T. Imielinski, A. Swami, "Mining Associations Rules between Sets of Items in Large Databases," In: Proc. of the ACM Conference on Management of Data, 1993, pp. 207–216.

[2] R. Agrawal, R. Srikant, "Mining sequential patterns", In Proc. Int. Conf. on Data Engineering, pp. 3-14, 1995.

[3] C.W. Cheng, G.S. Martin, P.Y. Wu, M. D. Wang. "PHARM – Association Rule Mining for Predictive Health," IFMBE Proceedings, 42, pp. 114–117, 2014.

[4] D. Dua, E. Karra Taniskidou, "UCI Machine Learning Repository," Irvine, CA: University of California, School of Information and Computer Science, 2017.

[5] S. Kalia, M. Greiver, X. Zhao et al. "Would you like to add a weight after this blood pressure, doctor? Discovery of potentially actionable associations between the provision of multiple screens in primary care." J Eval Clin Pract.24: pp. 423–430, 2018.

[6] B. Kamsu-Foguem, F. Rigal, F. Mauget, "Mining association rules for the quality improvement of the production process," Expert Systems with Applications, Vol. 40(4), pp. 1034-1045, 2013.

[7] R. Khade, N. Patel, J. Lin, "Supervised Dynamic and Adaptive Discretization for Rule Mining," SDM Workshop on Big Data and Stream Analytics, 2015.

[8] H. Liu, F. Hussain, C.L. Tan, M. Dash, "Discretization: An Enabling Technique," Data Mining and Knowledge Discovery, vol. 6(4), pp. 393–423, 2002.

[9] A. Y. Rodríguez-González, F. Lezama, C. A. Iglesias-Alvarez, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, E. M. de Cote, "Closed frequent similar pattern mining: Reducing the number of frequent similar patterns without information loss," Expert Systems with Applications, 96: pp. 271-283, 2018.

[10] A. Y. Rodríguez-González, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, J. Ruiz-Shulcloper, "Mining frequent patterns and association rules using similarities," Expert Systems with Applications, Vol 40:17, pp. 6823-6836, 2013.

[11] G. E. Schwarz, "Estimating the dimension of a model," Annals of Statistics, vol. 6 (2), pp. 461–464, 1978.

[12] E. Suzuki, J. M. Żytkow, "Unified algorithm for undirected discovery of exception rules." Int. J. Intell. Syst., 20: pp. 673-691, 2005.

[13] M.Y. Shih, J.W. Jheng, L.F. Lai, "A Two-Step Method for Clustering Mixed Categroical and Numeric Data," Tamkang Journal of Science and Engineering, vol. 13(1), pp. 11–19, 2010.

[14] S.C. Tan, P.S. Lau, "Time Series Clustering: A Superior Alternative for Market Basket Analysis," In: Proc. of The First International Conference on Advanced Data and Information Engineering, pp 241–248, 2013.

[15] S.J. Yen, C.K. Wang, Y. S. Lee, "Mining Consumption Association Rules," Journal of Information Science & Engineering, Vol. 32:2, p271-285, 2016.

Air Quality Monitoring in Mauritius

Praveer Towakel, Shivanand P. Guness, Amar Seeam and Girish Bekaroo
Middlesex University (Mauritius)
Uniciti, Flic en Flac
Mauritius
Email: {p.towakel,s.guness,a.seeam,g.bekaroo} at mdx.ac.mu

Xavier A. Bellekens
University of Abertay
Dundee
Scotland
Email: {x.bellekens} at abertay.ac.uk

***Abstract*—The aim of this project is to develop a low cost air quality monitoring system. The proposed system used the Raspberry Pi board, Arduino board, Grove sensors and Microsoft's Azure based cloud service for data storage and analysis. The data was captured from the 10th of May to the 31st of August 2017 in Bonne Terre, Vacoas. It was found that during the data capture period there was one occurrence in May and two occurrences in August where the $PM_{2.5}$ and PM_{10} were above the 25 ug/m3 and 50 ug/m3 level based on WHO guidelines.**

***Index Terms*—Air Quality, Azure, Sensor, MQTT**

I. Introduction

The project aims at investigating the ambient air quality in Mauritius. The proposed system will use Internet of Things (IoT) devices equipped with gas sensors to capture the level of pollutant in the air at various locations around the island. The monitoring devices consist of low cost devices such as Raspberry Pi 3 and Arduino boards. The main pollutants being monitored are Ozone (O_3), Carbon Monoxide (CO), Oxides of Nitrogen (NO_x) and Particulate Matter ($PM_{2.5}$ and PM_{10}). Additional environmental factors such as temperature, atmospheric pressure and relative humidity are also measured. The sensor data along with information such as the identifier of the monitoring system, time-stamp for the data captured, GPS coordinates of the sensing location are captured at regular interval during the data and sent to the Microsoft Azure platform.

II. Literature Review

This is the literature review for air quality monitoring. The review will look at the different studies carries out, the method used, the devices and the outcome of the study.

A. Air Quality Monitoring System

In recent years a number of research has been conducted on air quality monitoring systems. Air Quality Monitoring systems can be either fixed or portable and it is either to be used indoors or outdoors. The pollutants being monitored in indoor environment are different from the ones being monitored in the outside environment [1], [2].

1) Indoors: Chen et al. [3] proposed a system to monitor the Indoor Air Quality (IAQ) of Microsoft offices in China. The system consisted of using a Dylos 1700 [4] monitoring system on each floor of the building. The Dylos is used to collect $PM_{2.5}$ and Particulate Matter 10 (PM_{10}) every minute from each floor and stored to a local server. Every ten minutes, the average of the Particulate matter (PM)s readings are calculated and sent to the cloud. The cloud platform collects data from all the different offices. It also collects data such as the outdoor concentration of $PM_{2.5}$ and meteorological data such as wind speed, humidity, temperature and atmospheric pressure hourly. The data is used to evaluate the performance of the Heating, Ventilation and Air Conditioning (HVAC) system in filtering out the outdoor $PM_{2.5}$. The system was also used to highlight the HVAC that were not working and might require repair such as the changing of filter. The data is also used to predict the amount of time required by the HVAC systems to bring the indoor IAQ to safe levels by using an Artificial Neural Network (ANN). The system was 87% accurate to predict the time required to purity the indoor air to acceptable level and also decreased the time required and thus saving both time and energy. In Kim et al. [5], an indoor air quality monitoring system particulate matter was proposed. The system consisted of an Apple iPod Touch, an Arduino board and a Dylos DC1100 [6] device. The Dylos DC1100 [6] was used the measure the level of PM in the air at a 15 second interval. The Arduino was used to retrieve the data from the Dylos [6] device, convert it into a modulated tone and transfer it to the IPod via the audio jack. The data was sent to the inAir server and shared data was received using WiFi. The display on the iPod was refreshed every 5 minutes.

2) Outdoors: In Jovasevic-Stojanovic et al. [7], two methods were proposed to measure the personal exposure of an individual - direct and the indirect methods. The direct method involved providing the participant with a set of sensors consisting of NO, NO_2 and O_3 electrochemical sensors from Alphasense [8] and an Android application call ExpoApp. The Android application collected both the data from the sensors via Bluetooth and the data from the accelerometer on the phone to determine the physical activity of the participant, to send to the CitiSense servers. In the indirect method, air quality data obtained from the portable devices, along with data from fixed air quality monitoring units were used to create a map for each air pollutant. The fixed monitoring units consisted of 25 AQMesh [9] and 9 EB700 units with similar gas sensors as the portable unit and an optical sensor to monitor PM. If sufficient data was not available, the Land Use Regression model was used to extrapolate the air pollution data to generate air quality data for the whole city. Nikzad et al. [10], conducted a study with sixteen commuters in San

Diego using the portable sensor and using the mobile app from the CitiSense project. The study showed how the data collected by crowd sourcing system can provided a higher resolution and help to identify areas of pollution that regional monitoring station would not be able to spot. These systems also provide the user with a better idea on their daily exposure to pollutants. Following the study conducted in [7], Jovasevic-Stojanovic et al. conducted another study [11] on the four years the Citi-Sense project ran. It was found that the EB700 units, the CO sensor and the Dylos monitors for PM had a good correlation with their respective reference instruments. Both CO_2 and NO_2 sensor had a good correlation in [7] but had lower values in [11]. The AQMesh [9] had lower correlation with the reference instruments for the different sensors and the score were lower than the EB700 units. As the the Atmospheric units, the NO_2 had lower correlation compared to the other NO_2 sensors in the [9] and EB700 platforms. The PM monitors in this unit were not working.

In Aguiar et al. [12], proposed a low cost air quality monitoring system for the city of Brazilla in Brazil. The monitoring system consisted of a CO/VOC(MiCS-5521) sensor, a temperature sensor and a humidity sensor. Data was collected from different locations in Brazilia including the main bus terminal. The sensor board enabled data to be retrieved via Bluetooth and USB or sent wirelessly over either mobile or WiFi networks. In Brienza et al. [13] a low cost real-time cooperative air quality monitoring system called uSense was proposed. The uSense system was designed to be a low cost, low energy, wireless sensor nodes. It relied on the use of gas sensors such as CO(TGS-2442),NO_2(MiCS-2714) and O_3(MiCS-2614) sensors and long lasting batteries to capture air quality data. The data collected was send to the uSense servers over the internet using WiFi. The uSense system was installed by individuals on their property and they were free to choose to share their data with the community. In Alvear et al. [14], proposed to use an Unmanned Aerial Vehicle (UAV) to monitor the air quality. They opted for an open multicopter design i.e. all the components of the UAV should be off-the-shelf, open-sourced and modifiable. A Raspberry Pi [15] was used to control the UAV. The sensors used were the CO_2, CO, O_3,$PM_{2.5}$, PM_{10} and the air quality sensor and they were mounted on a GrovePi [16] shield. The GrovePi shield was mounted on the Raspberry Pi.

In Northcross et al. [17], an existing PM monitoring system from Dylos was modified and evaluated against a DustTrak 8520 [18] and an E-bam [19] from Met-One which is a beta attenuation monitor. The evaluations were conducted both with air samples containing known pollutant concentrations in the lab and in ambient air. The modifications done two intermediate bins were added t the particle bin size, increased data logging with better time resolution and the battery was extended. The modified devices was tested in the laboratory using Ammonium Sulphide, Polystyrene Latex Sphere (PLS) of $0.5\mu\mathrm{m}$ and particulates generated by burning of wood. Both the DuskTrack and the modified Dylos monitor were adjusted against the E-bam system before testing in ambient air. The daily mean of the modified device was within 80% of the two other systems.

III. System Overview

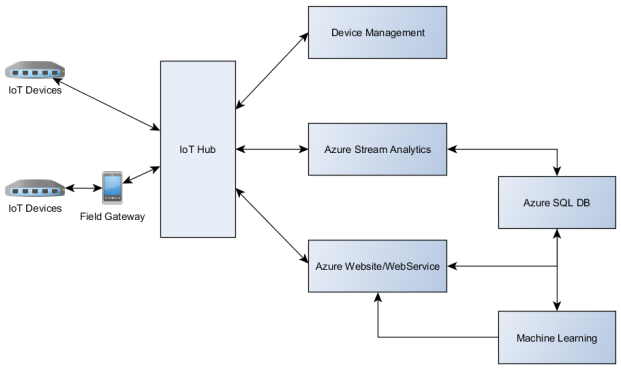


Fig. 1. System Overview

As it can be seen in Figure 1, the different proposed system will be using Microsoft's Azure [20] Cloud services to manage the data collected.

A. Hardware

1) Board: The Raspberry Pi 3 Model B [15] board includes Wireless/Bluetooth, one ethernet port and four USB 2.0 ports. It has a Quad Core 1.2 Ghz Broadcom CPU with 1 GB of RAM. An Arduino with a Grove shield was also used to connect the Grove based sensors.

2) Sensors: The sensors selected are Grove sensors and USB sensors. The Grove sensors can be easily attached to the board via a Grove shield provided and no soldering is required. Because no soldering is required, the maintenance of the of the air monitor system easy. And similarly the USB sensors can be attached to the USB ports of the Raspberry Pi. Most sensors have a lifespan of about two years and can get damaged i.e. the membrane can get poisoned by pollutants, and thus would have to be replaced.

a) Particulate Matter(PM_{10} and $PM_{2.5}$) Sensor: The Nova SDS011 [21] is used to measure the PM in the air. In contrast to other PM sensors such as the PPD42NS [22] from Shinyei which uses a heater to heat and the air and cause it rise into the chamber with a laser diode, the SDS011 sensor uses a built-in fan move the air to be sampled. The scattering effect caused when the beam from the laser diode falls on the particulates is used to determine the size and the amount of PM. The Nova sensor has been used in a number of studies [23]–[25] and this is why it was selected for this study.

b) Dust Particle Sensor: The air quality sensor [26] is a low cost sensor and it is responsive to a large selection of harmful gases such as carbon monoxide, formaldehyde , acetone etc. The sensor can be used in an indoor environment and if it is exposed to highly polluted air for a long period of time, there might be a decrease in sensitivity of the sensor. The sensor has quick response and a low power consumption.

The sensor used is a MP503 which is a metal oxide based sensor where the conductivity of the sensor changes with the concentration of a gas i.e.˜the higher the concentration of the gas, the higher is the conductivity of the sensor. This sensor can detect alcohol from 50 ppm to 1000 ppm [26].

B. Software

The system was implemented using Node.JS [27] on the Raspberry Pi. The Node.JS script was used to poll the data from the different sensors at regular intervals - in our case every 5 minutes. The data collected was converted to JSON format and sent to Azure.

C. Proposed Systems

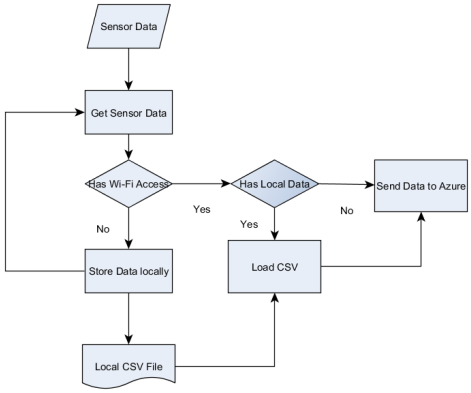


Fig. 2. Managing sensor data

a) Set up: The fixed air monitoring system consists of the Raspberry Pi board and both the Carbon Monoxide/Air Quality sensor via the Grove shield and the Nova SDS011 PM sensor. This system logs the data for $PM_{2.5}$, PM_{10} and CO for the location. Thus, the system can be used to calculate the Air Quality Index (AQI) level based on the data of the three pollutants.

b) Power: The system can use power from the mains. The system could include a battery as a backup in case of blackouts or power failure.

c) Connectivity: The data collected by the system could be relayed in real time to a server over Wi-Fi or saved to the local storage in the format of a comma-separated values (CSV) file. Being a fixed system, the system could be connected to the home network and the Wi-Fi connection could be used most of the time. The local storage could be used in situation where wireless connectivity is lost. The backup CSV file could them be unloaded either automatically or manually to the server.

1) Cloud Deployment:

a) IoT Hub: The IoT Hub [28] is used to provide an access point for the different air quality monitoring systems to send their data. Bidirectional communication i.e.˜device-to-cloud and cloud-to device, communication is possible with the IoT Hub. The IoT Hub manages the identity of the different devices and it also allows for revocable access control. It also offers a secure communication pipeline into Azure using Message Queue protocols such as Message Queue Telemetry Transport (MQTT) [29] and Advance Message Queuing Protocol (AMQP) [30].

b) Stream Analytics Services: The Stream Analytics Service [31] is a real time event processing engine. It can accept different types of data stream such as the data being sent the IoT Hub as input and based on a predefined function perform ongoing analysis on the data. The output from the service can be plugged in to different services such as PowerBi or stored in SQL Databases hosted on Azure.

c) SQL Database: The SQL Database is used to store the data being processed by the Stream Analytics Service. The database will be the back-end of the website used to disseminate the air pollution data sent by the different sensing units.

d) Azure Web Application Services: The Web Application Service is used to host application to disseminate the air quality for Mauritius. Web services can also be hosted on the service and used to connect the data with mobile applications.

IV. Results

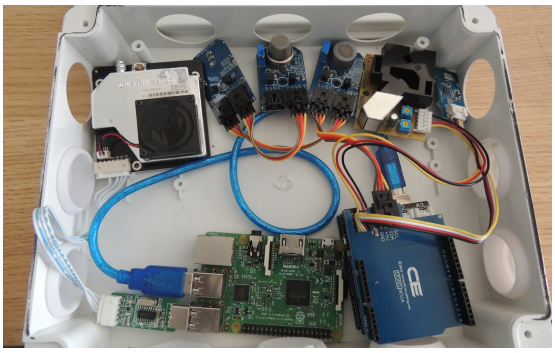


Fig. 3. Air Monitoring System

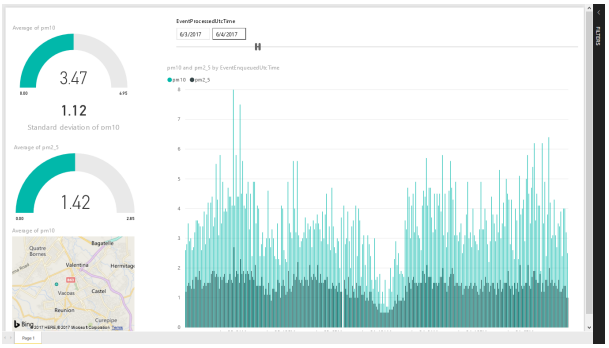


Fig. 4. PowerBI Report

In Figure 4, the report generated from the data captured is displayed. The dashboard is interactive so users can select the period to view the the data captured from the sensor. The location of the sensor is shown on a map and the mean and standard deviation of both $PM_{2.5}$ and PM_{10} is displayed.

The data for this experience was captured from the 10th of May 2017 to the 31th of August 2017 at the old Middlesex

TABLE I
SUMMARY OF $PM_{2.5}$ DATA

Months	Min	Max	Mean	StdDev
May	0.40	51.50	2.24	2.51
Jun	0.40	27.00	2.11	1.23
Jul	0.40	27.00	4.00	3.47
Aug	0.60	80.80	3.24	2.90
Mean	0.40	80.80	3.04	2.81

TABLE II
SUMMARY OF PM_{10} DATA

Months	Min	Max	Mean	StdDev
May	0.60	56.50	4.88	3.10
Jun	0.50	41.40	4.86	2.84
Jul	0.60	29.30	7.16	4.02
Aug	0.80	90.00	8.18	5.00
Mean	0.50	90.00	6.55	4.21

University(Mauritius) Campus which was situated at Avenue Droopnath Ramphul,Bonne Terre,Vacoas in Mauritius. .

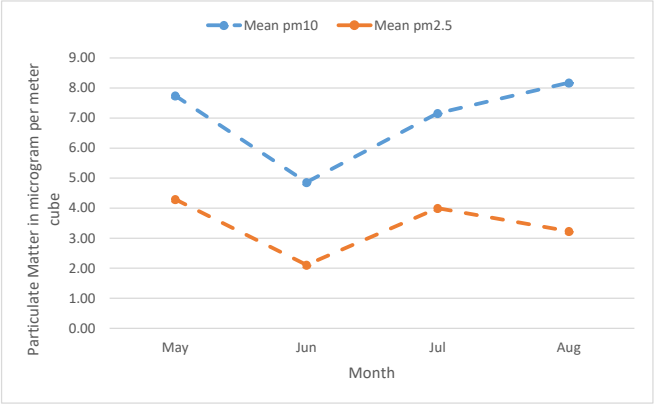


Fig. 5. Monthly mean of PM_{10} and $PM_{2.5}$

Based on the results in Figure 5, it can be seen that the mean $PM_{2.5}$ and PM_{10} levels did not exceed 9 micrograms per meter cube. The average $PM_{2.5}$ and PM_{10} measured were 6.85 ug/m3 and 3.26 ug/m3 respectively over the period the data was captured. Based on the data, it can be seen that the month of May and August had the highest levels of both PM_{10} and $PM_{2.5}$, and June was the month with the lowest levels of both PM_{10} and $PM_{2.5}$. In May, there was one occurrence one the 18th where the maximum PM_{10} and $PM_{2.5}$ were respectively 56.50 ug/m3 and 51.50 ug/m3. Similarly, in August there were two occurrences on the 23rd where the maximum PM_{10} and $PM_{2.5}$ levels were 67.93 ug/m3 and 75.07 ug/m3. On both days the limit was above the recommended guidelines for the health level set by the World Health Organisation (WHO) [2].

Figure 6, represents the mean hourly levels of PM_{10} and $PM_{2.5}$ detected. It can be seen that there is a peak for PM_{10} at 1 am in the morning and another peak around 5 am to 7 am in the morning. This peak could be due to people leaving for work and school children going to school. The level of both PM_{10} and $PM_{2.5}$ is fairly stable during the day except for

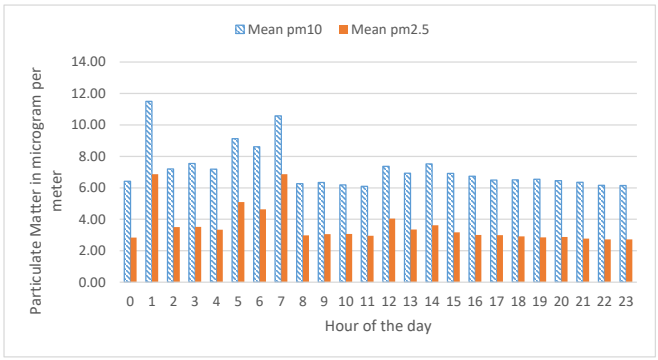


Fig. 6. Hourly mean of PM_{10} and $PM_{2.5}$

another slight peak from 11 am to 3 pm. This midday peak could correspond to people going for lunch and also the end of the school day.

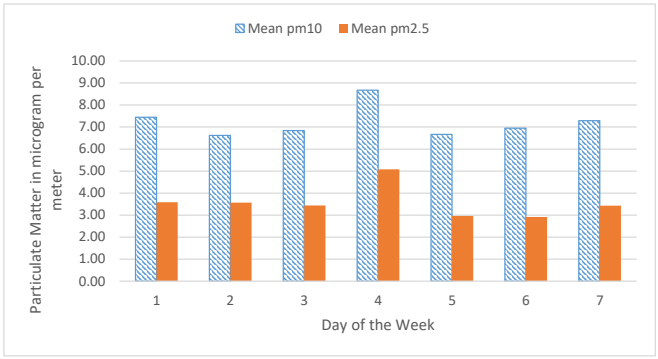


Fig. 7. Daily mean of PM_{10} and $PM_{2.5}$

The mean levels of PM_{10} and $PM_{2.5}$ based on the day of the week is shown in Figure 7. The day of the week is represented by numbers from 1 to 7, where number 1 represents Sunday and number 7 represents Saturday. As it can be seen, the levels of PM_{10} and $PM_{2.5}$ is fairly stable over the week except for Wednesday where there is a peak.

V. CONCLUSION

This project was an initial proof of concept to validate the equipment i.e. the low cost sensors we used along with Azure. Overall, the monitoring unit functioned well but due to a number of power cuts and outage which occurred during May to August 2017, some data were missing. The advantage of using a cloud based system such as Azure helped us prevent out data from being corrupted during the power outages.

As for the proposed future works, we are going to analyse the sensor data further along with the atmospheric pressure, humidity and temperature data we have collected. The next iteration of the air quality monitoring system would be solar based to make use of the sunny weather in Mauritius and also to prevent our data capture and sensors to be affected by power outage.

ACKNOWLEDGMENT

We would like to extend our thanks to the Mauritius Research Council who provided us the grant to make this project

possible and also to Microsoft for providing us the opportunity to use Azure for 1 year via their Azure for Research program.

REFERENCES

[1] D. Kotzias, K. Koistinen, S. Kephalopoulos, C. Schlitt, P. Carrer, M. Maroni, M. Jantunen, C. Cochet, S. Kirchner, T. Lindvall, J. McLaughlin, L. Mølhave, E. de Oliveira Fernandes, and B. Seifert, "The INDEX project - critical appraisal of the setting and implementation of indoor exposure limits in the EU," Tech. Rep., 2005. [Online]. Available: http://www.szu.cz/uploads/documents/chzp/ovzdusi/Vnitrni_ovzdusi/index_jrc_2005.pdf

[2] WHO, "Air quality guidelines. global update 2005. particulate matter, ozone, nitrogen dioxide and sulfur dioxide," Tech. Rep., 2006. [Online]. Available: http://www.euro.who.int/en/health-topics/environment-and-health/Housing-and-health/publications/pre-2009/air-quality-guidelines.-global-update-2005.-particulate-matter,-ozone,-nitrogen-dioxide-and-sulfur-dioxide

[3] X. Chen, Y. Zheng, Y. Chen, Q. Jin, W. Sun, E. Chang, and W.-Y. Ma, "Indoor air quality monitoring system for smart buildings," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '14. New York, NY, USA: ACM, 2014, pp. 471–475. [Online]. Available: http://doi.acm.org/10.1145/2632048.2632103

[4] Dylos Corporation, "DC1700 BATTERY OPERATED AQM," 2010. [Online]. Available: http://www.dylosproducts.com/dc1700.html

[5] S. Kim and E. Paulos, "InAir: Sharing indoor air quality measurements and visualizations," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '10. Atlanta, Georgia, USA: ACM, 2010, pp. 1861–1870. [Online]. Available: http://doi.acm.org/10.1145/1753326.1753605

[6] Dylos Corporation. (2008) Dylos DC1100. [Online]. Available: https://www.sylvane.com/media/documents/products/dylos-dc-1100-laser-particle-counter-owner's-manual.pdf

[7] M. Jovašević-Stojanović, A. Bartonova, D. Topalović, M. Davidović, and P. Schneider, "An implementation of citizen observatory tools used in the CITI-SENSE project for air quality studies in belgrade," 2016.

[8] Alphasense. Alphasense. http://www.alphasense.com/index.php/air/products/. Accessed: 2017-1-10. [Online]. Available: http://www.alphasense.com/index.php/air/products/

[9] "AQMesh." [Online]. Available: http://www.aqmesh.com/product/

[10] N. Nikzad, N. Verma, C. Ziftci, E. Bales, N. Quick, P. Zappi, K. Patrick, S. Dasgupta, I. Krueger, T. . Rosing, and W. G. Griswold, "CitiSense: Improving geospatial environmental assessment of air quality using a wireless personal exposure monitoring system," in *Proceedings of the Conference on Wireless Health*, ser. WH '12. San Diego, California: ACM, 2012, pp. 11:1–11:8. [Online]. Available: http://doi.acm.org/10.1145/2448096.2448107

[11] M. Jovašević-Stojanović, D. Topalović, M. Davidović, I. Lazovic, M. Živković, Z. Ristovski, and A. Bartonova, "Four years' experience of application of Low-Cost sensors in belgrade in the framework of the CITI-SENSE project," in *Proceedings*, 2016, pp. 46–49.

[12] E. F. K. Aguiar, H. L. Roig, L. H. Mancini, and E. N. C. B. d. Carvalho, "Low-Cost sensors calibration for monitoring air quality in the federal District—Brazil," *Journal of environmental protection*, vol. 06, no. 02, pp. 173–189, 2015. [Online]. Available: http://dx.doi.org/10.4236/jep.2015.62019

[13] S. Brienza, A. Galli, G. Anastasi, and P. Bruschi, "A low-cost sensing system for cooperative air quality monitoring in urban areas," *Sensors*, vol. 15, no. 6, pp. 12 242–12 259, 26 May 2015. [Online]. Available: http://dx.doi.org/10.3390/s150612242

[14] O. Alvear, C. T. Calafate, E. Hernández, J.-C. Cano, and P. Manzoni, "Mobile pollution data sensing using UAVs," in *Proceedings of the 13th International Conference on Advances in Mobile Computing and Multimedia*, ser. MoMM 2015. Brussels, Belgium: ACM, 2015, pp. 393–397. [Online]. Available: http://doi.acm.org.ezproxy.mdx.ac.uk/10.1145/2837126.2843842

[15] Raspberry Pi Foundation, "Raspberry Pi 3 with bluetooth & WiFi," 2016. [Online]. Available: http://uk.farnell.com/raspberry-pi/raspberrypi3-modb-1gb/sbc-raspberry-pi-3-mod-b-1gb-ram/dp/2525225

[16] "GrovePi+," 2016. [Online]. Available: https://www.seeedstudio.com/GrovePi%2B-p-2241.html

[17] A. L. Northcross, R. J. Edwards, M. A. Johnson, Z.-M. Wang, K. Zhu, T. Allen, and K. R. Smith, "A low-cost particle counter as a realtime fine-particle mass monitor," *Environmental science. Processes & impacts*, vol. 15, no. 2, pp. 433–439, 2013. [Online]. Available: http://dx.doi.org/10.1039/c2em30568b

[18] TSI, "Model 8520 DUSTTRAK aerosol monitor," 2002. [Online]. Available: http://www.tsi.com/uploadedFiles/_Site_Root/Products/Literature/Spec_Sheets/2980077_DustTrak_8520.pdf

[19] Met One, "E-Bam," 2008. [Online]. Available: http://www.metone.com/docs/ebam_datasheet.pdf

[20] Microsoft, "Microsoft Azure: Cloud computing platform & services," https://azure.microsoft.com/en-us/, accessed: 2017-4-11. [Online]. Available: https://azure.microsoft.com/en-us/

[21] "Laser PM2.5 Sensor-SDS011." [Online]. Available: http://www.inovafitness.com/en/a/chanpinzhongxin/95.html

[22] Shinyei Inc., "Dust particle sensor - PPD42NS," https://github.com/SeeedDocument/Grove_Dust_Sensor/raw/master/resource/Grove_-_Dust_sensor.pdf. [Online]. Available: https://github.com/SeeedDocument/Grove_Dust_Sensor/raw/master/resource/Grove_-_Dust_sensor.pdf

[23] K. N. Genikomsakis, N.-F. Galatoulas, P. I. Dallas, L. Candanedo Ibarra, D. Margaritis, and C. S. Ioakimidis, "Development and On-Field testing of Low-Cost portable system for monitoring PM2.5 concentrations," *Sensors*, vol. 18, no. 4, p. 1056, Apr. 2018. [Online]. Available: https://www.mdpi.com/1424-8220/18/4/1056/htm

[24] B. Fjukstad, N. Angelvik, M. W. Hauglann, J. S. Knutsen, M. Grønnesby, H. Gunhildrud, and L. A. Bongo, "Low-Cost programmable air quality sensor kits in science education," in *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, ser. SIGCSE '18. New York, NY, USA: ACM, 2018, pp. 227–232. [Online]. Available: http://doi.acm.org/10.1145/3159450.3159569

[25] A. D. Schwarz, J. Meyer, and A. Dittler, "Opportunities for Low-Cost particulate matter sensors in filter emission measurements," *Chemical Engineering & Technology*, vol. 41, no. 9, pp. 1826–1832, Sep. 2018. [Online]. Available: http://doi.wiley.com/10.1002/ceat.201800209

[26] Seeed Technology Limited, "Grove - Air quality sensor v1.3 - Sensor - Seeed Studio," 2016. [Online]. Available: https://www.seeedstudio.com/Grove---Air-quality-sensor-v1.3-p-2439.html

[27] Node. js Foundation, "Node.js," https://nodejs.org/en/, accessed: 2018-8-31. [Online]. Available: https://nodejs.org/en/

[28] Microsoft, "Azure IoT hub overview," https://docs.microsoft.com/en-us/azure/iot-hub/iot-hub-what-is-iot-hub, accessed: 2017-4-7. [Online]. Available: https://docs.microsoft.com/en-us/azure/iot-hub/iot-hub-what-is-iot-hub

[29] ISO/IEC, *ISO/IEC 20922:2016 - Information technology –Message Queuing Telemetry Transport (MQTT) v3.1.1*, Std. 20 922, 2016. [Online]. Available: https://www.iso.org/standard/69466.html

[30] ——, *ISO/IEC 19464:2014 - Information technology – Advanced Message Queuing Protocol (AMQP) v1.0 specification*, Std. 19 464, 2014. [Online]. Available: https://www.iso.org/standard/64955.html

[31] Microsoft, "Introduction to stream analytics," https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-introduction, accessed: 2017-4-7. [Online]. Available: https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-introduction

WeChat®: The Present and the Future

Hamelin Charlotte
Skema Business School
Suzhou, China
charlotte-hamelin@outlook.com

Rondelli Solène
Skema Business School
Suzhou, China
solene.rondelli@skema.edu

Ivan Coste-Manière
Luxury & Fashion Management
SKEMA Business School
Suzhou, China
ivan.costemaniere@skema.edu

Farbos Laura
Skema Business School
Suzhou, China
laura.farbos@skema.edu

Malenge Emmanuelle
Skema Business School
Suzhou, China
emmanuelle.malenge@skema.edu

***Abstract*—The purpose of this article is to analyze the power of WeChat and question ourselves on the hypothetical expansion. We discovered WeChat by living in China. For us, it is an application which gather everything. It is Instagram, Facebook, Twitter in just one app. But since we are living in a world with so many differences and different laws, we want to know if this app can really encounter the same success in different countries. WeChat has a big place in the life of Chinese people and this is something which is very interesting. Everything is around this app. The payment method, the social life and the way of interacting with people in general. We decided to split our study in three parts in order to answer the problematic.**

Keywords—WeChatÒ, artificial intelligence, Facebook, regulation, intranet, rights, privacy, protection.

I. Introduction

Today, people can do almost everything with one device. But in China, you can do everything you want with one single app. This app is called WeChat (Weixin in China). WeChat was founded in 2011 and is owned by Tencent, an investment holding company principally involved in the provision of add-value services and online advertising services. Tencent is the first company in Asia to pass $500 billion market cap mark and currently surpass Facebook in market value. Since 1998, Tencent has invested in many companies and it is almost impossible to know the exact number. The company have first invested in social network, digital payment, entertainment, information and news, clouds providers and in artificial intelligence with Tencent AI Lab and Tencent Youtu Lab.

WeChat is one of the biggest mobile apps in China with more than one billion monthly active users (in the most recent reported quarter) and covers 94% of Chinese smartphone users. Have you ever imagine an app which gather Facebook, WhatsApp, Amazon, Candy Crush, Tinder and other functions? Well, WeChat does it.

WeChat functionalities are all free. It includes texting, calling friends, posting moments, paying bills, booking train, bus, flight tickets, renting a bike, making a doctor appointment, ordering food and even arranging a time slot to file your divorce papers with the relevant authority.

WeChat is not only used in the friend and family area, but also used for businesses. WeChat users can create different accounts.

Our study is organized in three parts. We will talk about the use of WeChat first, then in a second paragraph about WeChat vs US competitors and lastly about the data protection.

II. The use of WeChat

A. Figures

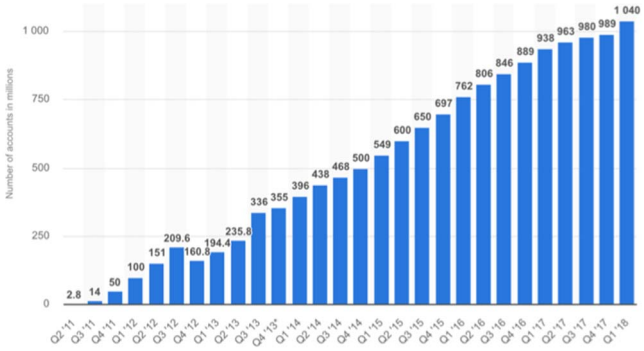


Fig. 1. No. of Wechat accounts in millions in the World (source : Statista)

WeChat application has an average of 902 million of daily active users, 38 billion of messages sent every day, 800 million users of WeChat Pay. Every day the payment transaction volume exceeds 600 million CNY. In 2018, WeChat was officially available in 25 countries and in 13 currencies. WeChat is currently available in 18 different languages.

B. Theory of contestable market

In China, the access to Google or YouTube is restricted, so WeChat represents a good alternative proposing several services in one application. Based on the theory of the contestable markets by the American economist William Baumol, WeChat is not a contestable market because barriers to entry exist and many applications do not work in China. In the US market, monopolies are not allowed with antitrust law because the country wants to protect the consumers and promote competition.

C. The additional features

Since 2011, the giant Tencent never stopped to improve WeChat application. They built a brilliant marketing ecosystem and became the new luxury digital trend in China. The app was initially used to send text messages or call friends, then became in a few years a super intranet app which gathers almost everything.

Moreover, WeChat does not only profit to personal users, but it also creates massive opportunities to businesses. Numerous companies in China have an official account and post regularly articles on it. They are able to communicate through WeChat and share information. They also can have their own online shops, mini-programs have been created to provide a new way of selling products and increase online consumers.

Furthermore, in the marketing prospective, WeChat creates opportunities for the companies to increase the brand awareness and the number of subscribers by using KOLs to talk or promote products. Meili United Group is a good example of WeChat mini program success, which has attracted more than 20 million new customers since 2016. Those mini programs could be an easy way for Chinese KOL to monetize and might be an unfair path to bleach money.

D. WechatPay

According to the Payment and Clearing Association of China, the number of transactions made through non-banking mobile apps from 2013 to 2016 increased from 3.8 billion to more than 97 billion. According to WeChatPay, ‘’WeChat has been certified to comply with the Payment Card Industry Data Security Standard (PCI-DSS)’’. Ma Huateng, chairman of Tencent, has undertaken many projects to develop the blockchain including the creation of a virtual money and the launch of an electronic invoice system. By creating a virtual money, Tencent is expecting Chinese consumers not only to invest on it as they did with the bitcoins, knowing that 60% of bitcoins investors are Chinese, but also to promote with the aid of several banks the investment on financial products. This system was developed to simplify the invoice process by removing all the intermediaries. It is also a way for the tax regulator to monitor and control the process of invoicing and tax collection. Tencent has recently partnered with Shenzhen government to battle tax evasion.

E. Artificial Intelligence

WeChat can be considered as an artificial intelligence app in many aspects. It has its own team of AI working constantly on developing speech recognition, natural language processing and computer Visio. For example, more than forty thousand medical institutions have allowed their patients to book and pay online, which allow Tencent to collect data in order to improve AI algorithms.

Thanks to a partnership between WeChat and Babylon, the users can have access to a virtual healthcare assistant. Tencent pushed the idea further by investing in iCarbonX : they develop a digital representation of individual to help perfecting personalized medicine. However, all these information are recorded and we could question ourselves on the use of these data.

III. WECHAT VS US COMPETITORS

A. The power of Wechat

To go a little bit further, we could easily say that WeChat could be the main world leader on the market. Indeed, Tencent has an incredible bargaining power, and became a nightmare for big companies such as Google, Facebook. Because it has enough money and strength to buy those companies. WeChat is a massive success which gathers everything. It succeeded by putting everything on simplicity. And in a world where everything goes fast, where companies can disappear in a day, massive companies are easily threatened. Something to notice is that, in Western countries for example, people are only using Google because they are used to it. They simply are not aware of WeChat functionalities and how it is used in China. But it is impossible to deny the power of this by looking at those numbers.

In 2006, the action of WeChat was valued at 3 HK dollars. In 2010, 30 HK dollars. And today, they are valued at 300 HK dollars. On the market, WeChat had an incredible rise. People who have invested on WeChat are expecting amazing returns on their investments. WeChat is estimated to be worth $100bn (Source: Forbes)

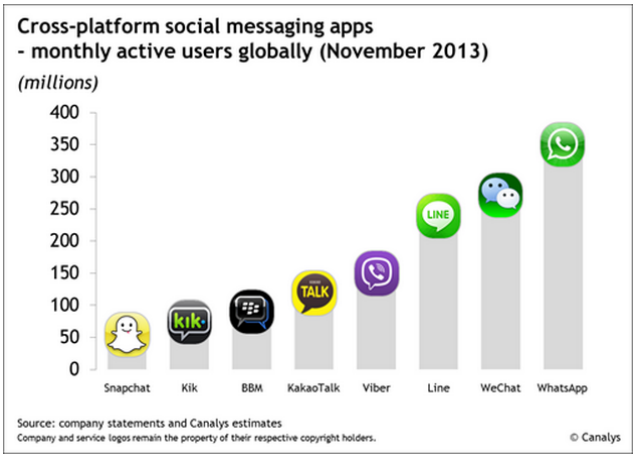


Fig. 2. Monthly active users globally of social media messaging apps (*company statements and Canalys estimates)*

B. Surpassing American companies

In November 2017, WeChat eclipsed Facebook with a market capitalization value of 529.9 billion, overtaking Facebook at 528.8 billion. The Tencent stock price has also doubled between 2016 and 2017. Again, the numbers can be explained by the high Chinese consumers acceptance towards new technology and due to the restricted access to other US companies such as Apple, Microsoft, Amazon.

We have to underline that Facebook will never reach the level of integration of WeChat. If Facebook or other US giant tech company want to reach this level of integration, they would have to issue national IDs (which means accepting a permanent control), convincing the bank industry.

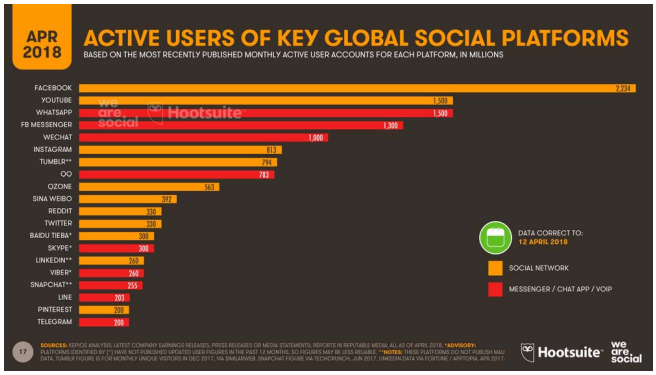


Fig. 3. Active users of key global social platforms (source : Wearesocial)

WeChat is not only a social app but it is an ergonomic and complete application. The "an-app-within-an-app" model is innovating and Facebook appears as an outdated platform. Even if, with Facebook, there are menus for customers so that they can communicate with companies accounts. In China, with WeChat potential, clients are well informed about the novelties and brands products. They can access directly to sales on the app. We might even wonder why Tencent, as big as it is, did not buy yet its American competitor.

C. *Leaders battle*

Let us take another example WhatsApp. One of WeChat best advantage's for Chinese is to communicate online with others knowing that more and more Western applications are forbidden by the government. We can compare the US leader communication application, WhatsApp, to the Chinese leader. Between the two applications, there are less than a 500 million user gap. This can be easily justified by the number of Chinese population but not only. First of all, according to China History, the increase of Chinese people connected to internet by using their cellphone came around 2010. Then, the precise location of these users is very important to detail knowing that WhatsApp can be used all around the world and WeChat was created for the Asian market. Nowadays, the number of daily users for both applications keeps increasing. For WhatsApp, foreigners in China either continue to use it or are becoming new users in order to keep contact with friends and family. For WeChat, the increase of its daily users is either due to the young generation getting a new cellphone, to the new foreign users who need the application to stay in contact with friends and family in China or need WeChat for business.

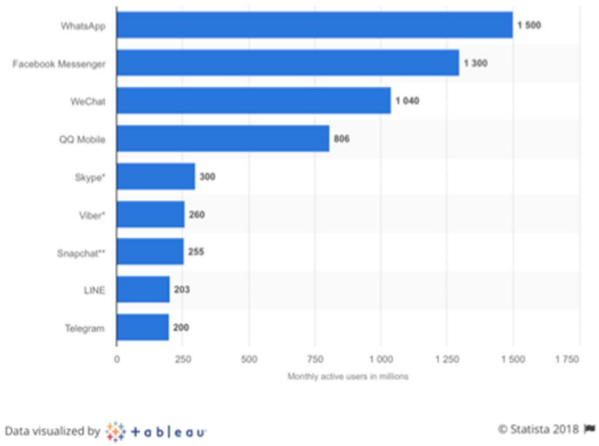


Fig. 4. Monthly active users in millions (source : Statista2018)

Fig. 5. Top 10 foreign nationalities in China (source: livinginBeijing)

D. *A successful future?*

Therefore, by comparing both new users, WeChat is getting more consumers in the West than WhatsApp is in China. Since China has opened its door to the world in 1980, Chinese and the West have never been that close in term of travels, new friends and businesses. Following the previous point, WhatsApp get new users in China thanks to the VPN. But, if the Chinese government decide to really reinforce the barrier and cut off the most famous VPN used by foreigners, such as Express VPN, the number of new users of WhatsApp will decrease while the number of WeChat users will considerably increase. To sum up, by comparing the market of the two applications, WeChat could conquer America and Europe … while WhatsApp cannot really become as popular in Asia. The market of WhatsApp is nearly saturated in opposition to WeChat.

As we described it, WeChat is able to surpass its main competitors. But we are not living in a homogeneous world. From a country to another, the laws and the basic rights are completely different. And this is the last point we need to mention in our study.

IV. DATA PROTECTION

A. *Implementation of WeChat outside China*

Even if WeChat is already used by more than 100 million people over 1000 outside of China, the application seems to have difficulties to be fully implemented.

The first example we can give is Europe. Compared to China, the protection of personal data is a fundamental right. And the problem with WeChat is that this platform does not respect this right. In 2017, some users of the app accused WeChat to share their personal data with the Chinese authorities. Indeed, in order to use the app, users have to accept the privacy policy of WeChat and it is explicitly saying that all information shared or watched are kept by WeChat to create personalize advertising, such as user's visiting WeChat companies, products bought or the time spend on the app. Legally speaking, implementing the use of WeChat like in China in Europe seems impossible right now. And, since May 2018, a new law protecting personal data use has been implemented. In the USA, the situation is also quite complicated. With the scandals around the NSA or more recently the Cambridge Analytica scandal, the United States have and might implement new laws and regulations to protect the data of the citizens.

B. *Data cloud issues*

In order to develop the use of WeChat around the world, we need to harmonize the legislation. This might be the most difficult point and one of the things which prevent WeChat to be used all around the world.

Some of the privacy policies is an obstacle for the international development of WeChat. Indeed, some personal information can be disclosed to the Chinese authorities if required. WeChat is allowed to activate microphone and camera, have access to the user pictures, track the localization and copy the data. Personal users and companies are using WeChat for advertising, which means that they share data

directly to the Chinese government. In many countries, this might be recognized as a human right violation.

In addition, administrators of WeChat groups are also responsible for the members comments which means that some content can be censored. The Chinese government has also implemented a "social-credit system" to rank their citizens according to their behaviors. This system will not only use outdoor cameras or check criminal report but also will collect all the information, physical and personal via WeChat.

C. *How could we protect ourselves then?*

Behind the iceberg, we also have the problem of fake goods. WeChat help companies to take advantage by making benefits out of it. Even if they did follow the Chinese regulation and shut down many businesses. China is still known as the leader of "the copy of the world". According to the "Committee Colbert", Asia is the continent where fake products of all kinds are made. The Committee has announced that China was the leader with more than 85% of the world counterfeit products. Therefore, WeChat could also be an issue: more counterfeit and a bigger conflict between Luxury brands and China.

V. CONCLUSION

WeChat is the Artificial Intelligence application that unified many of features and allow Chinese people to have a simple and fastest way of life. WeChat could benefit to the other parts of the world by becoming the first application to be connected in every continent. Nevertheless, we have to point at the issues: the law which present a lack of harmonization and the people who have to adapt themselves. Therefore, Tencent needs to be aware of those issues including data collection, competitor, regulations varying from country to another in order for WeChat to be fully used worldwide. Will it be able to make an agreement with other governments by creating a «win to win» deal? Which tools can be created in order to harmonize the laws? Will WeChat buy Google? Do we want WeChat to become the eyes of the world?

ACKNOWLEDGMENT

We would like to thank Soochow University for this knowledge and precious information that helped us to write this article.

REFERENCES

[1] Agarwal, Prateek. "Theory of Contestable Markets." *Intelligent Economist*, Intelligent Economist, 5 Jan. 2018, www.intelligenteconomist.com/theory-of-contestable-markets/.

[2] "Guide to Antitrust Laws." *Federal Trade Commission*, 15 Dec. 2017, www.ftc.gov/tips-advice/competition-guidance/guide-antitrust-laws.

[3] Baumol, W. J., Parzar, J. C., & Willig, R. D. *Contestable markets and the theory of industry structure*, 1988, San Diego, CA: Harcourt Brace Jovanovich.

[4] Sweney, Mark. "Tencent, the $500bn Chinese Tech Firm You May Never Have Heard Of." *The Guardian*, Guardian News and Media, 13 Jan. 2018, www.theguardian.com/business/2018/jan/13/tencent-the-500bn-chinese-tech-firm-you-may-never-have-heard-of.

[5] "Cross-Platform Social Messaging Apps Are Making Social Networking Increasingly Competitive." *Canalys*, www.canalys.com/newsroom/cross-platform-social-messaging-apps-are-making-social-networking-increasingly-competitive.

[6] Elodie. "WeChat Compte Désormais plus D'un Milliard D'utilisateurs à Travers Le Monde." *Journal Du Geek*, Journal Du Geek, 6 Mar. 2018, www.journaldugeek.com/2018/03/06/wechat-un-milliard-utilisateurs-monde/.

[7] "A Giant Cage." *The Economist*, The Economist Newspaper, 6 Apr. 2013, www.economist.com/special-report/2013/04/06/a-giant-cage.

[8] GlobalWebIndex. "GlobalWebIndex - Digital Analytics, Audience Insight Tools & Trends." *Latest Social Media Trends Report - GlobalWebIndex*, www.globalwebindex.com/.

[9] Kane. "Why Facebook Can't Be WeChat – Root Ventures – Medium." *Medium*, Augmenting Humanity, 9 Feb. 2018, medium.com/root-ventures/why-facebook-cant-be-wechat-3fdb14d70cf9.

[10] Revesz, Rachael. "A Company You've Never Heard of Just Eclipsed Facebook Last Week." *The Independent*, Independent Digital News and Media, 27 Nov. 2017, www.independent.co.uk/life-style/gadgets-and-tech/tencent-facebook-china-wechat-revenue-500-billion-alibaba-honour-of-kings-a8076861.html.

[11] Kosoff, Maya. "This Chinese Messaging App Is Taking the Country by Storm - and Facebook Should Pay Attention." *Business Insider*, Business Insider, 10 Aug. 2015, www.businessinsider.com/wechat-why-it-dominates-china-2015-8.

[12] "Facebook and WeChat, the Main Differences." *Marketing China*, 21 Feb. 2018, www.marketingtochina.com/facebook-wechat-main-differences/.

[13] Revesz, Rachael. "A Company You've Never Heard of Just Eclipsed Facebook Last Week." *The Independent*, Independent Digital News and Media, 27 Nov. 2017, www.independent.co.uk/life-style/gadgets-and-tech/tencent-facebook-china-wechat-revenue-500-billion-alibaba-honour-of-kings-a8076861.html.

[14] Bonifacic, Igor. "Amazon Reportedly Working on WeChat-like Messaging Service." *MobileSyrup*, 17 July 2017, mobilesyrup.com/2017/07/17/amazon-anytime-chat-app/.

Challenges and Opportunities of Artificial Intelligence in the Fashion World

Mariapaola Saponaro
SKEMA Business School
Suzhou, Jiangsu China
mpsaponaro@skema.edu

Diane Le Gal
SKEMA Business School
Suzhou, Jiangsu, China
diane.legal@skema.edu

Manjiao Gao
SKEMA Business School
Suzhou, Jiangsu, China
manjiao.gao@skema.edu

Matthieu Guisiano
SKEMA Business School
Suzhou, Jiangsu, China
matthieu.guisianodemarez@skema.edu

Ivan Coste Maniere
Director Luxury & Fashion Management
SKEMA Business School
ivan.costemaniere@skema.edu

***Abstract*—This article is depicting the Strengths and weaknesses of Artificial Intelligence related to the improvement of customer online and offline experience, and the possible methods in order to measure them. These methods include both researches non-based and based on interviews. The presence of AI in the retail industry is becoming a key component of the customer experience. Through a deep analysis of existing tools to extract information, we try to explain ways to interpret them, in order for companies to create a real usage out of them, either on online or offline retail experience. Hence, with this research, we also want to provide an insight on how this experience could be improved in the future, and how it will most likely be inherent to our daily customer experience.**

Keywords—Artificial Intelligence, fashion, customer experience, improvement, delivery gap, overspill, customer value

I. INTRODUCTION

Artificial Intelligence has been around since 1956 and has made giant leaps: beating the best human at chess, beating the best human at complex strategy game Go. Recently brands have started to adopt this technology for the core service. There is no doubt that the relationship between machine and human has changed. AI covers computing technology that can do things typically requiring human intelligence, from understanding language and recognizing visuals to making plans and solving problems.

Artificial Intelligence has been gaining lots of attention in the recent months, and is a growing technology used in different fields. Through the collection of organized data, AI and its algorithm has the ability to take out useful, meaningful information. Providing a full customer experience is the goal of every brand, and AI is more and more incorporated in it. It is believed that "By 2020, 85% of the customer interactions will be managed without human" – Gartner, and we can already notice that AI has been more than present in our past shopping experiences, either online or offline. The main objectives in its use for customer experience are to improve it, by time and efficiency, it allows us to answer easy questions that doesn't need to be asked anymore, just by the use of specific information. It also provides a more personalized experience to each customer, which makes them feel more special, as AI is speaking in their same language.

Thus, the difficulty to constantly improve customer experience has led to studies about how customers are experiencing a most interactive service, while the improvement of technology regarding clients' analyses have provoked a large amount of data quite uneasy to interpret. That is why the text mining method has emerged as way to take information from textual data, but it does not allow a very accurate analysis. However, when this method includes elements such as customer experience's important elements and service methodologies, it provides a more relevant data analysis. It allows companies to know the importance of interactive service processes on customer experience. It is also able to change the context of the study according to the time and to be used in different business domains, that is why it is called an open learning model.

However, when a brand integrates AI into the fabric of its core data, the information it is able to access will be much richer. While such a development then poses questions about what a brand does with that data, whether it is appropriate ethically and how marketers retain trust.

Due to the online competition, many e-stores give to customer the opportunity to personalize the pages so that they can use it more easily. The study about this help provided to customers is expensive because it requires several employees working on the systems. That is why the current researches were done to find new strategies by evaluating customer aid functions. Their behavior's study allows a simulation of the virtual shopping of users. Many experiments used a similar approach, and the results has shown that a personalized page is not always the best customer aid function for all the users in comparison with simpler ones.

II. AI IN ONLINE EXPERIENCE

AI has been mostly used on online platforms, such as online shopping malls. Chatbots are very well implemented, they can answer questions, and guide the customer through his online shopping experience, by helping him and even

recommending items that could be liked or matching the already purchased ones. This can be done thanks to the data collected during the time the customer was browsing the website, his past shopping experiences and matching other users' data. By doing so, the AI can provide a personalized experience to each customer, which is similar to offline shopping experience with an assistant but provided at distance.

Furthermore, Artificial Intelligence is closely linked with shopping applications using Augmented Reality as well as computer vision-driven image recognition and predictive inventory. Augmented Reality, or AR, can be defined by a virtual interface adding to the real-world complementary elements, such as texts, sounds, animations, and virtual objects that can be clothes. It works thanks to a camera, jointed to another device which can be a smartphone, glasses, a helmet or a screen. The real world and the virtual information are connected together through geo-localization and captors put inside the device, so that the user is well situated in the field he is standing in.

As an example of AR used in fashion retail, Alibaba is a multichannel application, that is to say that many brands items are sold through its virtual wardrobe, allowing customers to recall what they have tried on in-store while they also can read other buyers' or sellers' recommendations, so that customers are able to make one piece of clothes matching with other brands' ones. In few words, FashionAI innovation allows us to acquire a very wide sight of the trendy products in a well personalized way. In which way could this device be ameliorated?

In terms of pure fashion, it is a great innovation to become able to see yourself wearing new clothes and accessories you have selected on a smart screen. However, there is still a lack nowadays, because this vision does not allow customers to feel the texture of a cloth or the fabric it is made with, it is missing a more natural attribute to the experience. Thanks to the amelioration of the AR, the technology could become more and more precise in the future, so that the vision of the materials used in the confection of the chosen clothes could allow a more and more better appreciation of materials' texture even though it is not tried on, that is to say that it could improve fashion sales on e-commerce platforms, by providing a more in depth and real experience out of the point of sale.

III. AI IN OFFLINE EXPERIENCE

Several are the innovation technologies that stores are trying to introduce as a new way of customer experience. Nowadays FashionAI is growing as a system used to accurately observe and learn how to help the consumers in their style choices depending on what clothes they have tried in a store. One of these Fashion innovative devices consists in a screen able to recognize the clothes chosen and worn by a customer, and to suggest him or her other items matching with the tried one. Some of these screens are called smart mirrors, installed into stores' changing rooms. There use is mostly oriented in the help for consumers to find the articles in the boutique. This is called the customer experience improvement.

To give an example, Alibaba has signed a partnership with the Hong Kong Polytechnic University and Guess, which has launched a FashionAI store in July 2018. Customers have to apply with a Taobao code and their faces are scanned. The smart mirror then suggests them other purchases fitting with what they are trying, but also with their previous purchases. They choose the clothes or accessories they want to try, and they arrive quickly to them in the desired size and colour. They can also interact with the smart mirror if they request any change.

The customer experience improvement can also be seen with the personalization of the items that customers are about to buy. For example, in Paris' Faubourg-Saint-Honoré, the brand Berluti is using AR to allow its clients to choose any characteristics they want on their shoes, such as any kind of leather, which will be later confectioned by an artisan.

Another interesting case, Japan has been experimenting high-technologies hangers for years. These devices are capable to advice a certain combination between the chosen article and another from the same boutique. These hangers have been developed by teamLab after this company had noticed, through its e-commerce experience, that the articles worn by models — especially Japanese idols such as in Vanquish stores in Tokyo — were sold more easily that those which were just presented on the racks. Indeed, when a high-tech hanger is picked by a customer, a model wearing the cloth put on it with another item appears on a screen.
Thus, this kind of new devices lead to predictive fashion because it shows recommended and related items, so that more sales and profits are generated. Indeed, the information is then collected by the Artificial Intelligence so that the brand can have a deeper knowledge of the sales trends. To carry on giving the example of the high-tech hangers of Vanquish in Tokyo, their implementation had permitted to sales to double in 2011.

In addition to the smart mirrors and connected hangers, the Artificial Intelligence industry is thinking about creating avatars which would be copies of true persons. This sector could be worth $150 billion in 2030. The Hong Kong entrepreneur Adrian Cheng has made an investment of US$10 million in ObEN, a start-up using Artificial Intelligence technology for the creation of 3D smart avatars, whose appearance, voice and behaviour would be based on their true models. The prospective of this investment is to reach millions of new customers and create experience that will shape the future of retail. These avatars would be called PAIs, what means Personal Artificial Intelligence and by using them concierge and consumers can receive real-time information and personalised shopping advice. When the technology will be first available, mall visitors will be able to use tables provided by retailer, walk up to a piece of design and engage with an avatar representing a celebrity.

In other terms, although e-commerce is more and more developing, offline stores are still going to stay sustainable thanks to the new technologies' investments. However, we could be asking if this kind of technical innovations in offline retail will continue to preserve this movement to keep being modern and attractive. Indeed, despite technologies being a precious help for both customers and

retailers, the constant evolution endangers the offline area to be quickly out of fashion. So, what could be able to save this sector? The winning recipe may consist in the alliance of humanity and technology.

IV. EVOLUTION OF INNOVATION

According to Joëlle de Montgolfier, senior director of the strategy sector at Bain & Company, online purchases in the luxury industry are nowadays about 8%, and could increase to almost 25% until 2025, so that we have to review and renew the commercial approach. Similarly, it will be necessary to modernize more and more the boutiques to allow the story telling not to be lived only through a screen but also in offline stores.

In terms of retail in China and in USA the offline experience is becoming more and more declining. However, Alibaba, with its 15 stores across China, is going to put digital into the offline retail world. Indeed, Kanzhen Peng, the secretary of China Chain Store said: "In the age of mobile Internet, the merging of online and offline [retail] is a trend". In other words, Alibaba is going to open the border between online and offline in its stores, that is to say that customer will come to them to command items available in the storage. José Blanco, who is the chief executive of Guess in China, holds that it was necessary to have a partnership with Alibaba in order to keep innovating in retail, because the customers' habits have already evolved. "At Guess, we need to innovate in real time." Thus, Guess will be the first brand to digitalize stores thanks to this precious partnership, because not only smart mirrors will be developed, but also smart racks and smart entire fitting rooms in stores all around China.

However, the technological approach is not the only element having to be rethought and improved. Indeed, clients are desiring to be spellbound when they enter a luxury boutique, because luxury is not fulfilling a need but a dream. The human relationship is the primordial complement to the digital prowess of a brand. A fashion store has to be like a palace, or at least like a sweet home for the customers, and the digital is used to have a better knowledge of their needs thanks to data collect.

In 2016, Sony has come up with intelligent contact lenses capable of taking pictures. Even though this technology is still under research, we can totally imagine it to be applied in the fashion field. Since smart contact lenses have been developed with a camera embedded inside, it is more than possible that this technology will be applied in various fields. For the fashion field, it could equip the customers with the ability to identify the products they need efficiently therefore to boost a superior experience for customers. And on the other hand, for the retailers, this technology would be an efficient tool to locate the demands of inventories as well. Realizing this technology is not a matter of how but when.

V. MEASUREMENT OF CUSTOMER EXPERIENCE

It would be interesting thinking about organizations investments in design of consumption experience effects. Nowadays marketing investments are practiced without experiential measurement effects but are mainly based on credibility of the supplier, relationship skills and imitative behavior spread (managerial mode). Due to the lack in focusing on the main objectives, several power marketing projects fail in the realization. In other cases, it frequent to refers to indicators as sales trend and the number of visitors at an organized event but with no attention regarding relevant partial effects: number of cognitive effects, sales variation as proxy of behavioral effects. In relation to the world of state of art, Michela Addis suggests useful technics aimed to the result measurement of consumer experience:

Researches not necessarily based on consumer's interview

a) study of physiologic reactions of emotion at the peripheral level
b) study of physiologic reactions of emotions at the central level

In order to analyze experiences and their emotion dimension is it necessary use tools able to disclose individuals in their unconsciousness. Although physiologic components have been studied since 70s in terms of consumer behavior, knowledges on this topic are still few, because it is still difficult finding a match between physiologic answer and related emotion. In neurosciences field new investigation techniques have been developed which accurately disclose the relationships between physiological reactions, both at peripheral and central level, and individual stimuli. It is possible to say that these techniques represent the most important refence in terms of research methods development on the effects of experience. However, there are still adjustments to do, in fact, measurements do not are always accurate. Although these techniques are able to measure the emotion experienced, they do not give accurate indications in terms of the emotion nature and do not detect its intensity. Therefore, they only detect change from one emotional state to another.

Researches necessarily based on consumer's interview:

a) Extensive survey to detect emotions
b) Experiments with questionnaire on consumer experience

The objective of surveys is merely descriptive and the sample is necessarily large, while the experiments are focused on identify the links between variables cause and effects and for this reason sample could be smaller. Both methods use the questionnaires, needed in order to receive consumers answers. It can be affirmed that surveys constitute one of the most popular marketing method, in particularly, in detecting consumer emotions. The researcher submits a stimulus to the consumers and the questionnaire consists of a multiplicity of questions with items related to positive and negative emotions in relation to the consumer reaction with this stimulus. The data collected constitutes the basis on which researchers develop statistical analyzes as multidimensional scaling, factor analysis and cluster analysis. "The Costumer Experience Improvement Program" (CEIP), adopted by Microsoft, it is aimed to achieve a continuous improvement in the experience by the enlargement of information amount.

Although it is impossible to find a single model of reference, there are at least two shared and recognized emotions properties: the bipolarity, in which the emotions are placed in a continuum between two opposite semantic

poles (happy-sad) and the continuity, in which is possible to measure the intensity of the emotional phenomenon through the use of continuous scale. The use of this method is frequent in order to analyze the impact of communication messages, layout of sales points and exposure techniques on emotions. However, there are some limitations to the method, in fact it is inadequate to notice the bonds between different emotions and emotion and knowledge that together shape the consumer experience. Moreover, its application has an unrealistic approach since the method is based on the assumption that the phenomenon can be broken down into components of experience or individual emotions.

Experimental design is a widely applied method in the physical sciences and it is used to investigate the links between the variables that regulated consumption experience. The principal way how to use this method is to isolate the effects of the variables in which researchers are interested in and control all the others involved. In terms of behavior, traditional consumer research has focused on the process choice from where purchase decisions are generated. A recent state of the art method, theory and application has been provided by Singer in the study of conscious experience. Approaches that are comparable in conventional consumer research include techniques of though-generation, protocols of problem solving and procedures of ideation-reporting. Under an experiential perspective it remains to extend this work which is cognitively oriented toward the investigation of consumption experience in all its possible aspects.

These technics were elaborated in order to disclose individual's emotions and the measurement of consumer's experience effects seems to be the ideal context. It is important to highlight that these technics focus only on the emotional dimension, leaving aside any other relevant variable of the phenomenon of consumer experience.

According to Forrester (April 2016), 70% of businesses put the customer experience as one of their top priority. This shows that improving loyalty and advocacy combining with maintaining high standards in a competitive market is not an easy strategy. Furthermore, the ways how to measure the results of costumer experience are not really well known. This is because it does not exist a clear-cut questionnaire address to shop visitors asking them how satisfied they were during their customer journey. Even if it exists, how is possible to measure their happiness and satisfaction?
Individual customers are individual, for this reason they are different. Although many marketers advance the theory according to which even if customer experience is about emotions and individual parameters, there are many common points in the way people experience things, fulfil task and value their own experiences there is still complexity around that topic.

This complexity has been creating the "customer experience gap" or "delivery gap". A survey by Bain & Company shows that more than 93% of management teams claim to be customer focused. So, they assumed having found two fundamentals reasons for the gap. The first one is based on a business paradox: when businesses try to increase revenues per customer, actions like raising transaction fees, end up alienating the core group of buyers. The second reason is related to the fact that good relationships are hard to build. Understand what customers really want, keep the promises and maintain the right dialogue to ensure satisfaction according to customers' changing or increasing needs it is extremely difficult. Although innovation technologies are growing day by day, there is still the lack of a right methodology in this field. Do we have to hope in a future machine which will be able to read in people's mind?

VI. CONCLUSION

Data collection adds new parameters, such as how the consumers are shopping (what they look at, how they look at it, from where, etc.) that are now essential for the brand to know more about how to provide the best experience for the customer. These parameters can be used in order to understand more about fashion trends, in the process of analyzing customers' likings, how they match outfits, to the designing of the point of sales and how everything is arranged for the customer to be looking for what he or she likes, but also the sounds, the fragrance that can be used in store. It is a way to create a store more attractive to customers, a better tool to targeting the right market, and give a new, refreshing experience. It would be a revolutionary way to come up with new product designs, ideas and predict the future trends. In particular, it could be an effective way aimed to attract the big target of new generations, whose is growing up in a more and more digital world, incorporating AI and new technologies in their shopping experience. It could create more awareness of the brands and incorporate new brand experiences.

However, thinking about the possibility of improving customer experience and the business itself through Artificial Intelligent could it actually work? Reducing human relationships and the approach proper of human mankind could damage customer perception and delight when shopping. Reducing contact with people could lead to a lack of something that an artificial intelligence could not replace: advices from shop seller in relation to costumer's characteristic that data collected in a robot could not disclose and the involvement and attention that just another human being could transmit to another. Could we like AI more than other humans? We have to make sure AI remains a tool for humans, and not the opposite. An AI will never replace the expertise and knowledge of a human being, especially in the fashion world because it is mostly a matter of taste, which is subjective.

REFERENCES

[1] Alibaba official web page: https://www.alibabagroup.com/en/about/overview

[2] C. Boos (April 27, 2018). How Fashion Shop Should (and Shouldn't) Embrace Artificial Intelligence). *Business of Fashion Professional*. https://www.businessoffashion.com/articles/opinion/how-fashion-should-and-shouldnt-embrace-artificial-intelligence

[3] J. Bughin, M. Chui, and B. McCarthy (August 2017). How to make AI work for your business. *McKinsey & Company* https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/how-to-make-ai-work-for-your-business

[4] J. Chan (July 10, 2018). Guess to be First Global Brand to Digitise Stores with Alibaba's FashionAI. *Luxury Society*. https://www.luxurysociety.com/en/articles/2018/07/guess-be-first-global-brand-digitise-stores-alibabas-fashionai/

[5] K. Chitrakorn (January 9, 2018). Adrian Cheng's K11 Invests $10 Million in AI Start-Up. *Business of Fashion*.

https://www.businessoffashion.com/articles/news-bites/adrian-chengs-k11-invests-10-million-in-ai-start-up
[6] M. Dupuis and Vanessa Zocchetti (Novembert 11, 2017). Réalité augmentée, boutique-maison : le luxe invente le shopping du futur [Augmented Reality, boutique-maison: the luxury is inventing the future way of shopping]. *Le Figaro Madame.* *http://madame.lefigaro.fr/style/realite-augmentee-boutique-maison-le-luxe-invente-le-shopping-du-futur-071117-135111*
[7] C. Ferlay-Naudin (December 2017). L'intelligence artificielle va chambouler l'industrie du luxe [The Artificial Intelligence is going t Disrupt the Luxury Industry]. *Sawi* https://sawissd.blog/2017/12/22/__trashed/
[8] Gartner Customer 360 Summit 2011 (March 30 to April 1st, 2011). CRM Strategies and Technologies to Understand, Grow and Manage Customer Experiences, Gartner Summits.
[9] M. Liu (November 11, 2017). Artificial intelligence starts to revolutionise luxury industries. *Financial Times.* https://www.ft.com/content/1c2a6b24-a514-11e7-8d56-98a09be71849
[10] G. Magana (July 6, 2018). Alibaba is piloting fashion AI technology. *Business Insider*. https://www.businessinsider.com/alibaba-starts-using-fashion-ai-technology-2018-7
[11] H. Nakata (May 9, 2013). High-tech hangers offer shopping tips. *The Japan Times*. https://www.japantimes.co.jp/news/2013/05/09/business/high-tech-hangers-offer-shopping-tips/#.W4ciIpMzat-
[12] J. Rapp (January 15, 2018). These Hyperreal Avatars Could Change the Way Chinese People Shop. *Jing Daily*. https://jingdaily.com/six-ways-ai-will-revolutionize-chinas-luxury-market/
[13] The real experience of luxury begins with artificial intelligence (April 19, 2017). *Raconteur*. https://www.raconteur.net/sponsored/the-real-luxury-experience-begins-with-artificial-intelligence
[14] https://www.iscoop.eu/customerexperience/#A_matter_of_emotions_and_interaction_customer_experience_and_customer_engagement
[15] http://www2.bain.com/bainweb/pdfs/cms/hottopics/closingdeliverygap.pdf
[16] Babis Theodoulidis, Jamie Burton, Thorsten Gruber, Mohamed Zaki http://journals.sagepub.com/doi/abs/10.1177/1094670514524625
[17] PaolaCecchetta (2011) http://tesi.cab.unipd.it/32806/1/Cecchetto_Paola.pdf
[18] Hyung Jun Ahn https://www.sciencedirect.com/science/article/pii/S0020025510000022
[19] Y. Sun (November 13, 2017). Alibaba's AI Fashion Consultant Helps Achive Record-Setting Sales. *MIT Technology Review*. https://www.technologyreview.com/s/609452/alibabas-ai-fashion-consultant-helps-achieve-record-setting-sales/
[20] Michela Addis (2007) "Ad uso e consumo. Il marketing esperienziale per il manager".
[21] Peter Leppik (July 29, 2015) A/B Testing for Customer Experience https://www.vocalabs.com/blog/ab-testing-customer-experience
[22] Morris B. Holbrook, Elizabeth C. Hirschman. The Journal Of Consumer Research. Vol. 9, No. 2 (Sept. 1982), 132-140 https://s3.amazonaws.com/academia.edu.documents/32052289/Holbrook.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1537947392&Signature=Vj%2BTgvkr%2BQCL9SAwWzDK%2BpCCxr0%3D&response-content-disposition=inline%3B%20filename%3DHolbrook.pdf
[23] Elisabeth A. Buchanan, Erin E. Hvizsah http://journals.sagepub.com/doi/abs/10.1525/jer.2009.4.2.37
[24] Todd jaquith(May 2nd 2016). Sony's smart contact lenses can record what you see. https://l.facebook.com/l.php?u=https%3A%2F%2Ffuturism.com%2Fsonys-new-contact-lenses-let-record-store-everything-see&h=AT2IAMX2xXv_e-a50LJnbxlF3iP5DyGHHg63DS_TcqjyTnKpqv2WNSqSqPUdGcnSY1GHpDewfxCXbvtdhj1VGEqEhcdw9uXzUKIcktDC7AqyDhKf8YazB2ZgCOvpQjqx8BTSlNF

Towards the Development of an Interlink Protocol for Disease Surveillance Data Aggregation: A Namibian Context

Nikodemus Angula and Nomusa Dlodlo
Namibia University of Science and Technology
Windhoek, Namibia
chcangula@gmail.com, ndlodlo@gmail.com

***Abstract*—Wikipedia's definition of semantic interoperability is about two or more computer systems having the ability to be able to communicate and exchange data in a meaningful way, which means data exchanged from different source can be understood by the other computer systems. Protocols specify interactions between the communicating entities. In this research, a protocol is developed as an interpreter of disease surveillance data from heterogeneous health information systems (HIS) in Namibian public health institutions for data semantic interoperability. This enables the District Health Information System (DHIS-2) and other health information silo systems to exchange health data and information, and specifically disease surveillance data. The study has produced a new interlink protocol which acts as a converter of data coming from multiple silo systems. Therefore, the interlink protocol has the capability to aggregate disease surveillance data from different data sources. This interlink protocol is based on JSON format.**

Keywords—health information systems, disease surveillance, interlink protocol, data aggregation

I. INTRODUCTION

Interoperability is the ability of diverse systems to work together, i.e. interoperate [1]. Interoperability is also defined as: "The capability to communicate, execute programs, or transfer data among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units" [2]. Wikipedia's definition of semantic interoperability is about two or more computer systems having the ability to be able to communicate and exchange data in a meaningful way, which means data exchanged from different source can be understood by the other computer systems [3]. In addition, a protocol is the set of rules that governs the communication of two systems [4]. Moreover, the role of the protocols is to specify interactions between communicating entities [5]. In this case, a protocol was developed as layer for communication purposes and as converter to aggregate disease surveillance data in Namibian public health institutions to aggregate disease surveillance data in the Namibian health institutions [12].

The lack of access to readily-available aggregated disease surveillance data among health professionals in Namibia creates a gap when health professionals have to make decisions timeously [13]. As it stands, health administrators working in the 14 regional sub- offices communicate disease surveillance information to the head office in Windhoek through USB or manually through sending technicians with physical files to the regions for import into the system. The CDC supports the MOHSS conduct disease surveillance activities, monitor and evaluate such activities, research and maintain HIS [2].

Currently data is communicated to the CDC through the Directorate of Special Programmes that deals with HIV/AIDS, TB, malaria, and related diseases and through primary healthcare directorates in the different regions of the MOHSS. However, because of poor and inefficient communication of disease surveillance information, there is little or no data coming in from the regions for real-time decision-making in cases where outbreak of disease requires prompt action [2]. The methods used currently to communicate data from the regions to the CDC are emails or phone calls. These methods are not appropriate in cases of emergency in disease control.

Therefore, a real-time access through an integration of HIS is required System (DHIS-2) with other health information systems in order to communicate and exchange health related information and data among different health information systems in Namibian health institutions [1]. This interlink protocol is based on JSON object format. In the Namibia health sector environment there is no known protocol as it stands that governs heterogeneous health information system and the aggregation of disease surveillance data between the health institutions [7].

Therefore, the study proposes a solution to address the challenge of a lack of exchanging and communicating disease surveillance data. The aggregation of disease surveillance data will allow the health professionals such as doctors, nurses, and the disease surveillance office to access and communicate disease surveillance data in real time. Despite unstructured data from different heterogeneous health information systems, this study developed an interlink protocol that would allow disease surveillance data to be aggregated coming from different data sources [6].

The paper is structured as follows: Section 2 is the problem statement. Section 3 is related work. Section 4 is the proposed framework. Section 5 is an explanation of how

the system works. Section 6 is a case study of application of the interlink protocol. Section 7 is the demonstrator, Section 8 is the business benefits of the system and Section 9 is the conclusion.

II. PROBLEM STATEMENT

Currently in Namibia there is a lack of a mechanism for semantic interoperability to aggregate disease surveillance from multiple heterogeneous HIS. There are no known protocols for semantic interoperability that have been adopted by the MoHSS and customized to the Namibian health environment. Currently the CDC, MoHSS, and public health institutions in Namibia encounter a challenge of a lack of disease surveillance data aggregation from heterogeneous HIS which negatively impacts health service delivery when health decision making is to be implemented within the Namibian health domain [14]. The lack of disease surveillance data aggregation among health institutions in Namibia creates a gap when health professionals want to access real-time disease surveillance information. As it stands, health administrators working in the 14 regional sub offices communicate disease surveillance information to the head office in Windhoek through USB or manually through sending technicians with physical files to the regions for import into the system [15].

The CDC supports the MOHSS conduct disease surveillance activities, monitor and evaluate such activities, research and maintain HIS [2]. Currently data is communicated to the CDC through the Directorate of Special Programmes that deals with HIV/AIDS, TB, malaria, and related diseases and through primary healthcare directorates in the different regions of the MOHSS. However, because of poor and inefficient communication of disease surveillance information, there is little or no data coming in from the regions for real-time decision-making in cases where outbreak of disease requires prompt action [16]. The methods used currently to communicate data from the regions to the CDC are emails or phone calls. These methods are not appropriate in cases of emergency in disease control. Therefore, a real-time access through an integration of HIS is required.

III. RELATED WORKS

The following is work or related systems that enable semantic interoperability.

A. The openEHR Archetypes Approach

According to [1] the openEHR archetypes approach is an approach that enables syntactic interoperability and semantic interoperability to allow for the exchange of data from multiple information systems. The purpose of this approach is that it enables semantic interoperability among sources and receivers [10]. This approach is not favourable for this research simply because of the uncoordinated efforts internationally and across health professions to match or aggregate heterogeneous systems currently. Further development of domain knowledge on archetypes is required.

B. Ontologies

Accordingly, the Enterprise Knowledge [8] states that ontology is model that is capable of organizing structured and unstructured information through organisations, their properties and the way they relate to each other and organizes the data in the same format. In other words, ontology consist of classes and properties. In addition, concepts are structured in a taxonomy and relations are set as non-taxonomical connections [10]. This approach is not favourable for this research simply because it doesn't clearly indicate how data is matched or aggregated among heterogeneous systems and because of the multiple interpretations of the same data.

C. Model-Driven Engineering (MDE)

According to [9], model-driven engineering technologies are capable of offering an approach which is promising in order to address the inability of third-generation languages to alleviate the complexity of platforms and express domain concepts effectively. This approach would also be lacking in enabling semantic interoperability since this research is an aggregation of data and not looking in depth at hardware or software.

D. Formal Specification of Context Interchange

Based on [10], context interchange is one of the strategies that represents a novel perspective which mediates data access where there are semantic conflicts among heterogeneous systems. This approach is not favourable simply because this is a developing research area and does not clearly indicate how data is matched or aggregated among heterogeneous systems.

E. XML

According to [13], XML is a set of rules for ensuring that encoding documents in a format that can be both human-readable and machine-readable.

- XML is extensible.
- XML allows the creation of self- descriptive tags or language that suits an application.

These features of XML can be adopted for semantic operability save for the fact that it is heavy as compared to JSON [11].

F. JSON

JSON or JavaScript Object Notation is a lightweight text-based open standard designed for human-readable data interchange [12]. JSON has the following characteristics:

- It is used while writing JavaScript-based applications that includes browser extensions and websites.
- JSON format is used for serializing and transmitting structured data over network connections.
- It is primarily used to transmit data between a server and web applications.
- Web services and APIs use JSON format to provide public data. It can be used with modern programming languages.
- JSON is easy to read and write.
- It is a lightweight text-based interchange format.

JSON is very useful when developing a web application where fast, compact and convenient serialisation of data is required. However, its flexible nature is the very thing that makes it less suitable than XML for transferring data between separate systems or storing data that will be read by third parties.

- JSON is more compact and can be easily loaded in JavaScript.
- XML is stricter and has support for schemas and namespaces**.**

The lightweight feature of JSON and the fact that it is language-independent would be of benefit in this research for development of the protocol**.**

IV. THE PROPOSED FRAMEWORK

When data is captured in different public health institutions, it is in different formats, meaning that each time a patient visits a different health institution, they have to open a new file. To enable interoperability of data, the application of an interlink protocol to disease surveillance data from silo health systems in Namibia is characterized by the following, as shown in Figure 1.

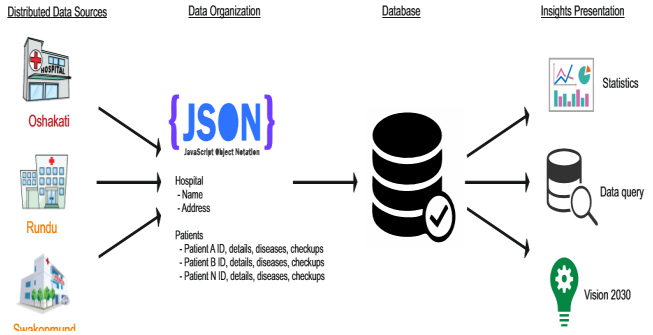


Fig. 1. Interlink protocol

There are JSON files from different public health institutions. As long as these files are in JSON format, the system will accept and produce statistics form them. Based on the information uploaded on the system, health professionals can view analysed data on the dashboard and can easily make quick decisions based on the information. The standard interface to enable aggregation of data is the rule of aggregating data is coded as a layer to allow remote systems and DHIS-2 system to exchange disease surveillance data with each other in a meaningful manner.

The aggregated data is used in the production of, for example, reports to be used in decision-making. The study therefore has identified a new protocol called interlink protocol in the Namibian public health sector that would act as converter to interpret data from different data sources.
An example of the code for interlink protocol in JSON is as shown below in Figure 2.


```
{"source":{"name":"Grootfontein Hospital",
        "id": "0020",
        "type":"state hospital",
        "physical_address":"",
        "postal_address":"",
        "telephone":"",
        "fax":"064-344",
        "email":"",
        "town":"Windhoek",
        "region":"otjozondjupa"
        },

 "records":[
        {"category":"cholera",
        "id":"960619727272",
        "name":"Jeremiah Fonseca",
        "sex":"m",
        "dob":"19 June 1996",
        "telephone":"0813536",
        "town":"Grootfontein",
        "region":"Otjozondjupa",
        "email":"",
        "date":"12 January 2016",
        "description": "Patient tested negative for cholera"
        }
        ]
}
```


Fig. 2. JSON object

This example protocol describes how the interlink protocol works. A user in the public hospital, CDC or MoHSS can organize their disease surveillance data in JSON format. Different health institutions in Namibia can organize their data as long as is in JSON format it can be organized regardless of its data sources. That format is the JSON object. This is an example of a JSON object for information generated from Grootfontein Hospital. It has records for different patients including the header with details of the particular hospital. This JSON object is in a standard format and so are the JSON objects for the other public health institutions.

V. HOW THE SYSTEM WORKS?

Different clinics/hospitals, community health workers (CHW) and community members login onto their systems, be they desktop or mobile-based and capture disease. The crowd-sourced disease surveillance data from the community members is captured via their cell-phones. The CHW capture disease surveillance data via their mobile devices. The systems automatically generate JSON objects.

This is further explained in Figure 3.

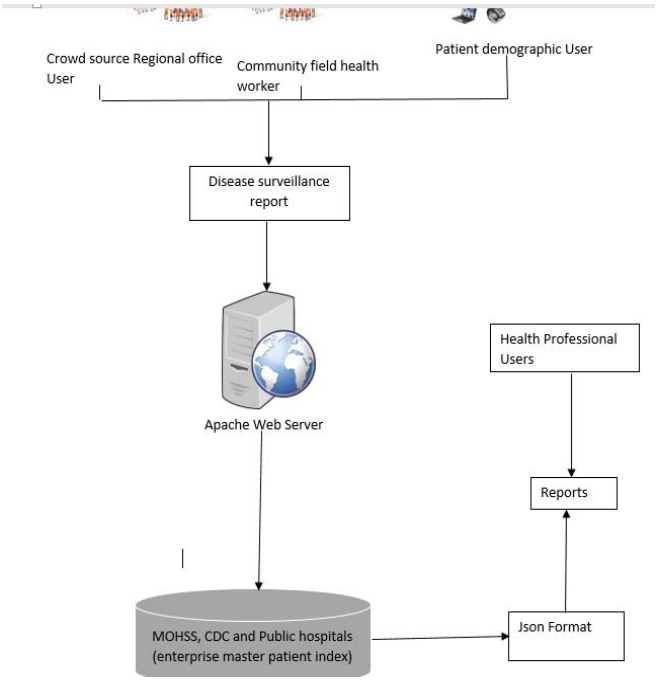


Fig. 3. Architecture of the system

Earlier we mentioned that the different HIS are heterogeneous in the sense that the data generated is not standard across all of them. Therefore, the JSON object is what creates this uniformity in the data, as shown in Fig. 4.

The crowd-source user, community health work, and the patient demographic user can gather disease surveillance data using devices like laptops, cell phones, and other electronic devices from any region in Namibia anywhere and at any time and report disease surveillance data in real time. The Apache Web server in Figure 3 initiates the connection between the database and the web application. This database is the enterprise master patient index (EMPI) at the MOHSS that stores data from different heterogeneous health information system provided that the data is in JSON format. The health professional users will be able to view reports of statistics such as charts, graphs etc.

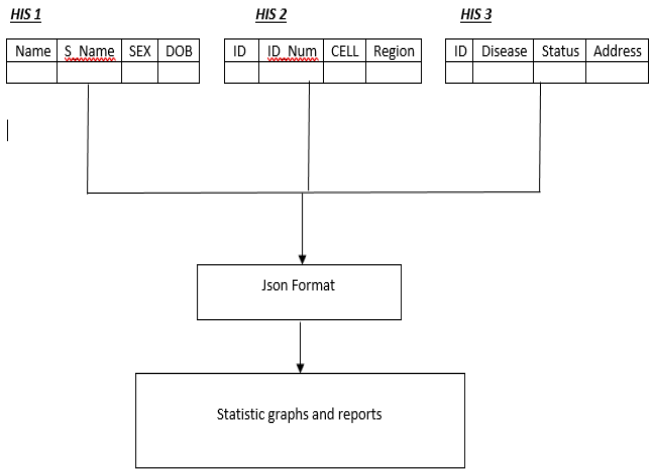


Fig. 4: Uniform JSON interface

V. A CASE STUDY OF APPLICATION OF THE INTERLINK PROTOCOL

The MoHSS operates silo HIS in Namibia's 14 regions in addition to a regional District health information system (DHIS-2) for each region and the main DHIS-2 in the MoHSS. The MoHSS also works closely with the Centre for Disease Control (CDC) in disease surveillance. Therefore, the need for all these distributed HIS to communicate and share disease-surveillance information in real-time. Therefore, integration of disease surveillance data through this Interlink protocol would mitigate that current problem facing the entity such as the delay in making health decisions and access to health information from all the 14 regions in Namibia.

Figure 5 is the use-case diagram. The crowd-source user, community health worker and patient demographic user can login to the system and submit reports and the system will generate JSON file automatically and integrate data from multiple heterogeneous health information systems which can be monitored and administered by the systems administrator at the head office in Windhoek. This administrator can be at the MoHSS and CDC. At the same time both users can terminate or logout from the systems.

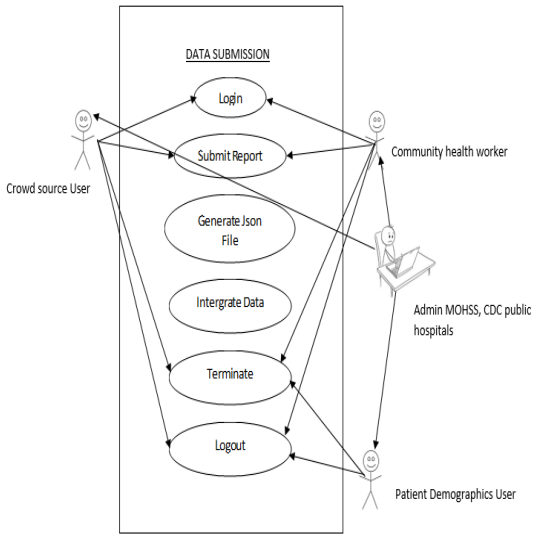


Fig. 5. Use case diagram

VI. THE DEMONSTRATOR

The disease surveillance demonstrator that utilizes this protocol is as shown below in Figure 6. The user can login through user login interface using their credentials, that is, the username and password.

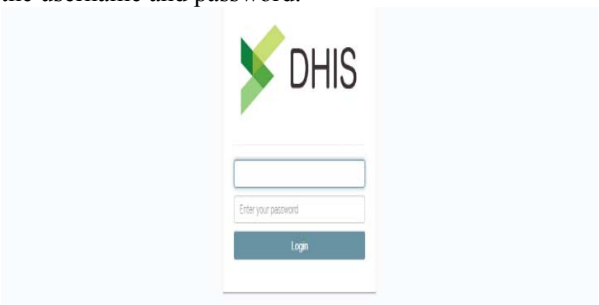


Fig. 6. Login interface

The user can login and view a dashboard with information uploaded from different data sources in all the 14 regions in Namibia, as shown in Figure 7.

Fig. 7. Distribution of diseases

The map symbolizes the geographical location a particular disease occurred in any of the 14 regions in Namibia, as shown in Figure 8.

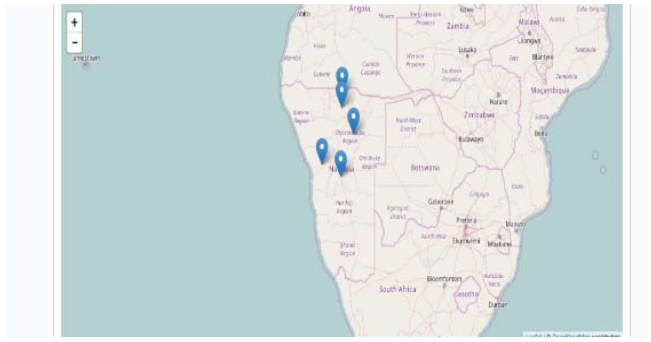


Fig. 8. Geographic location

The next screen enables a user to filter the type of disease a particular patient is suffering from. This is shown in Fig. 9.

Fig. 9. Filtering by type of disease

Figure 10 is a list of all patients in the integrated system and Figure 11 is a list of diseases. With the system manage tab the user can view patients, diseases, hospitals and patient records on the system. With the upload data tab, the user can upload data under this tab provided the data is structured into JSON format. With the system setting tab the user can create user account on the systems and administrator can manage the system under is tab.

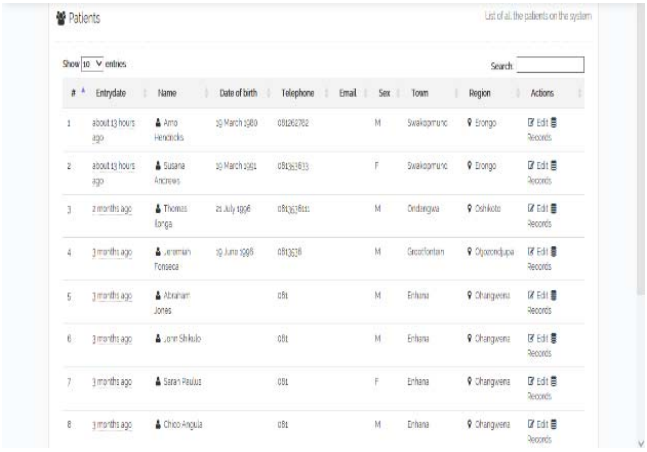


Fig. 10. List of patients on the system

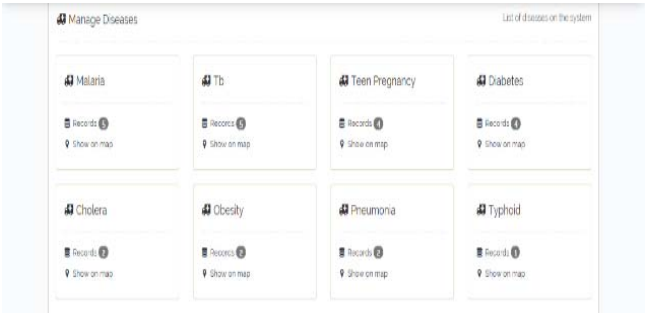


Fig. 11. List of diseases in the system

VII. BENEFITS OF THE SYSTEM

There are a number of benefits that the semantic interoperability of HIS achieves [17]. One of these is that the patient can access health services at any health institution and their medical history can easily be retrieved even if the record was not captured in the same institution they are visiting[18]. The doctors have a complete view of the patient's medical history irrespective of where they are. The system is able to aggregate disease surveillance data from multiple heterogeneous HIS in the Namibia health domain at the same time [19]. The system data exchange scheme and standards are able to permit disease surveillance data to be shared across clinicians, laboratory, hospital, pharmacy and patients regardless of the application or application vendor within the Namibian health domain. The system has a capability to allow the disease surveillance data exchange between health organisations in Namibia across organisational boundaries in order to advance the effective delivery of healthcare to individuals and communities. The Interlink protocol enables the aggregation of disease surveillance data from the different HIS which have different data structures and are developed using different software on varied platforms [20]. When data is automatically transferred from one system to the next, this reduces the occurrence of errors. The integration of disease surveillance data facilitates easy, informed and real-time decision-making. Therefore, the integration of data breaks down the silos in medical health. At the end of the day the customer, who is the communities has increased satisfaction.

The study is recommending future researchers to consider all aspects that were not covered in this research study. In addition, anything that was not included in this research study should be included in future research study. The interlink protocol should be further refined.

VIII. CONCLUSION

In the Namibian health domain there is no known protocol to govern or aggregate disease surveillance data from remote heterogeneous health information systems. Therefore, the study developed an interlink protocol that can aggregate disease surveillance data from remote systems. The protocol is based on JSON format. To test the protocol, a prototype was that uses the protocol was designed. The prototype enables the integration of disease surveillance data and the production of reports for decision-making. The health professionals in Namibia would use the system for fast decision making simply because they are accessing disease surveillance data in real time.

ACKNOWLEDGMENTS

We wish to thank the Namibia University of Science & Technology for providing facilities to undertake this study.

REFERENCES

[1] S. Garde, P. Knaup, E. Hovenga and S. Heard, "Towards semantic interoperability for electronic helath records," in Ehealth beyond the horizon: get it there: Proceedings of MIE2008, European Federation for Medical information 2008, 2008.

[2] A. Tatnall, Ed., Encyclopedia of portal technology and applications, IGI Amazon, 2007, p. 881.

[3] "Semantic interoperability," [Online]. Available: https://en.wikipedia.org/wiki/Semantic_interoperability. [Accessed 25 September 2018].

[4] B. Victoria, "Signal processing," in IEEE Pacific Rim Conference on Communications, Computers and Signal processing, 1987.

[5] W. Kusniercyk, "Nontological enginnering," in Proceedings of the International Conference on Formal Ontology in Information Systems, Amsterdam, 2006.

[6] D. Kalra and B. Blobel, "Semantic interoperability of EHR systems," Study in health technology and informatics, pp. 127-231, 2007.

[7] N. Angula and N. Dlodlo, "Towards a frameowrk to enable semantic interoperability of data in heterogeneous health information systems in Namibian public hospitals," in ICITS 2018, Ecuador, 2018.

[8] J. Diez, What is enterprise ontology?, Berlin Heidelberg: Springer, 2006.

[9] D. Schmidt, "Guest editor's introduction: model-driven engineering," Computer, vol. 39, no. 2, pp. 25-31, 2006.

[10] C. Goh, S. Bresson, S. Madnick and M. Siegel, "Context interchange: new features and formalisms for the intelligent integration of information," ACM Transactions of Information Systems (TOIS), vol. 17, no. 3, pp. 270-293, 1999.

[11] A. Ouksel and A. Sheth, "Semantic interoperability in global information systems".

[12] N. Angula, and N. Dlodlo, "Towards a framework to enable semantic interoperability of data in heterogeneous health information systems in Namibian public hospitals", Springer, vol. cham, pp. 835-845, Jan.2018.

[13] U. Batra, S.Sachdeva and S.Mukherjee,"Implementing healthcare interoperability utilising SOA and data interchange agent.Health policy and technology, "IEEE", vol.doi-4,pp.241- 255,April..2015.

[14] L. Guijarro, "Interoperability frameworks and enterprise architectures in e-government initiatives in europe and United States, Government information Quarterly", Elservier, vol.doi-24, pp.89-101, May.2007.

[15] O. Iroju, A. Soriyan, I. Gambo, and J. Olaleke, "Interoperability in healthcare: Benefits, challenges and Resolutions International of and applied studies", Innovative space of scientific research journal",vol. 3, pp. 262-270, May 2013.

[16] N. Schuurman, "Social perspectives on semantic interoperability: constrants on geograpical knowledge from a data perspective. Cartograhic: The international Journal for geographic information and Geovisualisation,", utpress, vol. catro-40, 47-61, November.2005.

[17] E.L., Siegel, and D.S. Channin, "Intergrating the healthcare enterprise:a prrimier:part1. Introduction.Radiographics," The technical framework", vol. b-21, pp.1339-134, February.2001.

[18] P.Sinha, G.Sunder,P.Bendale,M.Mantri and A.Dande " Integrating the healthcare enterprise (IHE) electronic health record:Standards, Coding Systems, frameworks, and Infrastructurs,"IHE" ,vol. pp.175-185, 2012.

[19] H. Van der Veer, and A. Wiles, "Archieving technical interoperability European Telecommunications Standards Institute,"European Telecommunications standards institute", vol. April.2008.

[20] K. Writting, "Health information exchange: Integrating the healthcare enterprise (IHE). In introduction to Nursing informatics, "Springer London" vol. pp.79-96, 2015.

A Training Utility for Estimating the Bowling Speed of a Cricketer Using Accelerometer Data.

Bhekisisa Nyoni
Department of Electronic Engineering
National University of Science and Technology (NUST)
Bulawayo, Zimbabwe
bhekiburgs@gmail.com or
bhekisisa.nyoni@nust.ac.zw

Magripa Nleya
Department of Electronic Engineering
National University of Science and Technology (NUST)
Bulawayo, Zimbabwe
magripa.nleya@nust.ac.zw

Busiso Mtunzi
Department of Electronic Engineering
National University of Science and Technology (NUST)
Bulawayo, Zimbabwe
busiso.mtunzi@nust.ac.zw

***Abstract*— This prototype system known as the Wearable Instantaneous Ball Speed Estimator (WIBASE) was designed to measure the bowling speed of a cricketer during training. When fast bowlers are training, coaches have to assess their ability to bowl consistently fast balls even when they are required to perform long bowling spells, hence the need for reliable, accessible and affordable equipment for measuring their bowling speed cannot be over emphasised. The WIBASE seeks to fill in this gap. It is made up of two hardware components; a computer and a wrist-worn electronic board that houses among other components, a 3-dimensional (3D) acceleration sensor. The system tracks the three-axis acceleration generated by the movement of the arm when delivering the ball and stores these values. The raw sensor data from three different sensors namely accelerometer, gyroscope and magnetometer is processed by a Digital Motion Processor (DMP) on the board in a process known as Sensor Fusion before it is sent via Bluetooth to the computer. The computer runs a Python script that receives the filtered acceleration which consists of both static acceleration and dynamic acceleration. The acceleration is numerically integrated over a minute period of time around the release point using the Trapezoidal method of integrating numerical data to derive the speed of the cricket bowler. The results obtained from the three sets of experiments that were conducted show that the WIBASE can track the 3D acceleration of the hand when bowling, derive the speed of the bowlers and display the speed on a computer while logging all the data into a file.**

Keywords*— *accelerometer, inertial sensors, sensor fusion, inertial measurement unit; motion analysis, numerical integration, cricket, training, bowling speed

I. INTRODUCTION

While there is a documented need for Zimbabwean cricketers to improve their fast bowling performance [1], research revealed that there is no equipment in Zimbabwe's cricket grounds for measuring the speed of cricket bowlers during training in order to improve their performance. The Radar guns that are traditionally used to measure the speed of bowling in cricket are only used here in Zimbabwe when there are international matches and the cost of hiring them together with broadcast equipment is said to be steep. This presents two challenges; exorbitant cost and unavailability. Video-based high-speed motion capture systems cannot be used either because they are expensive, time consuming and often require skilled operators for both data collection and data analysis [2].

Despite the unavailability of equipment for measuring the speed of a cricket bowler, the assessment criteria of bowlers during training sessions includes among other considerations determination of consistency and secondly ensuring that maximum bowling speeds are achieved. Both require the measurement of the speed of bowling.

Dadashi et al. [3] stress that monitoring performance is a crucial task for elite sports during both training and competition. This therefore means that there is a need for systems that can quantitatively keep track of performance. Mirabella et al. [4] discussed two types of tools based on image processing and the use of inertial sensors for continuously extracting and storing parameters of movement with a degree of accuracy. Qaisar et al. [5] concurs that inertial sensors are being used in numerous sports applications. They put forward that accelerometers were being used to measure and classify activity and effort levels in sports. A considerable number of studies [6] have been conducted on the application of Inertial Measurement Units (IMUs) in various sports. An IMU is defined as a combination of an accelerometer and gyroscope primarily.

Kevin and Darryl designed and constructed a wristband that measured the throwing speed of a football player. In their design, the picocontroller (PIC) calculated the speed of the ball by measuring the player's arm acceleration as the football was thrown. Their device was found to be accurate to within 10% of the velocity measurements obtained from a radar gun [7].

Since the research sought to perform numerical integration on acceleration data, the analysis of the methods of integrating acceleration data by Slifka [8] was an invaluable contribution to this research. The research took note of the crucial findings of Kajánek [9] and Sabatini et al. [10], to reduce errors in the estimate of speed from the integrals of acceleration. Analysis of the research by Spratford et al. helped in writing an algorithm for the release time [11].

With all this background study and analysis of findings, a training aid for fast cricket bowlers was developed and tested. Monitoring of fast bowlers using the Wearable Instantaneous Ball Speed Estimator (WIBASE) had the following advantages; freedom of motion since communication was wireless, 3 degrees-of-freedom tracking of acceleration, immediate feedback and data logging.

II. SYSTEM ARCHITECTURE

The WIBASE integrated hardware components and the complementary software for determining the speed of a cricket bowler and Fig. 1 shows the architecture of the system. The IMU required embedded software for it to work properly. While the laptop computer required a program

developed using Python for the IMU to work in tandem with it.

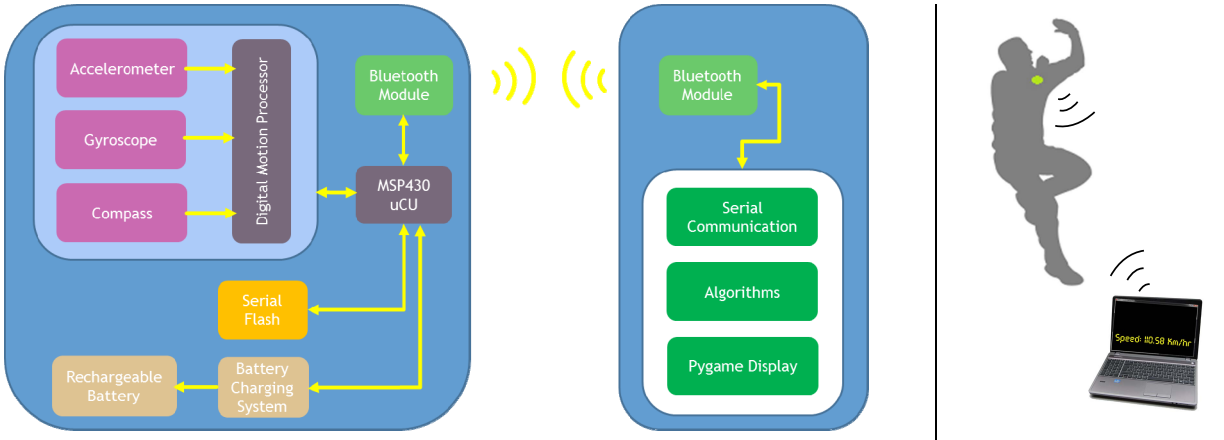


Fig. 1. The (a) Proposed System Architecture for the WIBASE (b) Conceptual view of the system in use

A. Hardware

The WIBASE was made up of two hardware elements; a wrist worn IMU and a laptop computer. Figure 2 depicts the IMU employed, known as the MotionFit SDK which is a product of TDK Invensense.

The core of this IMU is a miniaturized 4x4x1mm Motion Processing Unit (the MPU-9150) which contains the 3-dimensional (3D) gyroscope, accelerometer and magnetometer. The MPU also houses a DMP for implementing 9-axis MotionFusion algorithms. In addition to the MPU, the IMU also contains an MSP430 microcontroller, Bluetooth transceiver, 256Mbit serial flash for storing activity data and a rechargeable 3.7V, 110 mAh Li-Ion battery [12]. The IMU sits on a board that is 43.5mm x 37.3mm and weighs 36grammes. The ideal connection between the computer and the IMU was established by Bluetooth pairing. The rate at which the IMU sends sensor data to the second hardware element (computer) was determined from the update rate of the quaternions that it sends to the computer at a frequency of 50Hz [12].

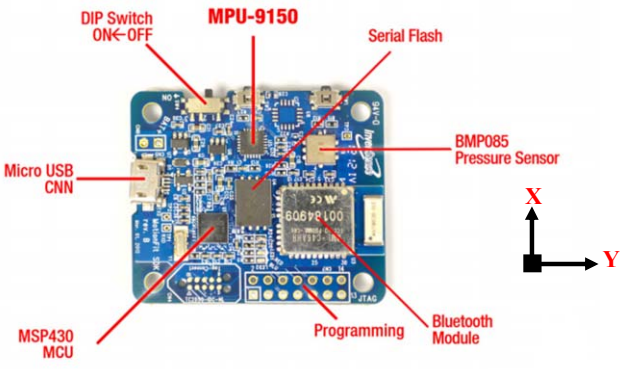


Fig. 2. The IMU known as the MotionFit Wireless SDK by TDK Invensense.

B. Software

Two software programs were developed for the system to function properly. The embedded software running in the IMU controls critical motion processing tasks including sensor device management, sensor bias calibration, sensor / motion fusion, and gesture detection. The embedded code was modified to increase the full-scale range of the system from ±2g to ±16g. This necessitated reprogramming the IMU.

Secondly, several methods were written in Python to develop the client program that ran on the laptop computer. The python script included different classes which enabled among other functions the ability for the laptop to establish a Bluetooth link with the IMU via a serial COM port and thus request the sensor data from the IMU. This program was designed to accept the sensors' data, process the received data further, calculate the speed and display the speed on a Graphical User Interface (GUI). A module of Python known as Pygame was used for implementing the GUI. Other modules that included the PySerial, Numpy, OS, Math, Scipy, MySQLdb, SYS, Time and Matplotlib were also used to implement various functionality. Countless modules that are available for the Python development environment render this language very powerful and user friendly, coupled with its simple syntax.

III. Methodology

Assumptions were incorporated into the design process of the system. The speed that is derived from the acceleration at the wrist of the bowling hand was approximated to the speed of the delivery since miniaturizing the sensor boards to sit on the fingertips was not possible in Zimbabwe. Secondly integrating acceleration over a short time interval was expected to deal with the problem of drift and give a good estimate of the speed. This is in line with the assertion by Yuan et al [13] who spelt out that the integrating method is not available in long duration velocity tracking. The third assumption was that the peak of the acceleration of the wrist was likely to occur at the release point. Finally, both the

static (gravity) and dynamic acceleration was used in the calculations since the experiments will be conducted under normal environmental conditions in which the force of gravity is always at play. The experimental procedure that was followed to collect the data and the blueprint for processing the data is discussed below.

A. Data Collection Procedure

Measurements were performed by three bowlers namely, the researcher, an amateur bowler and a professional bowler. The bowlers were asked to deliver 12 fast balls each while the system was recording the data of each delivery. Figure 3 shows the amateur bowler in action during the data capturing session in the nets at Bulawayo Athletic Club (BAC).

Fig. 3. Conducting tests with an amateur cricket bowler at BAC

The experiments were done in the nets because of the limitations in the distance covered by the Bluetooth wireless module on the IMU. Experiments were done to determine the extent or reach of the communication between the module and the computer. It was found to be effective within a radius of approximately 10 metres. This implied that the run up of the bowler up to the crease had to be within this range for the wireless communicate with the computer to persist.

Meanwhile the professional bowler was asked to bowl 12 deliveries as well while measurements were being logged into text files for analysis using MATLAB.

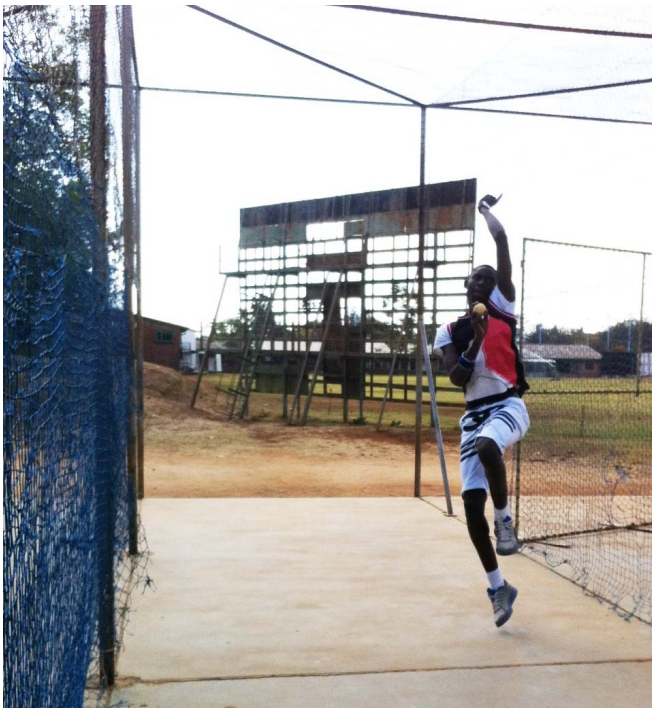


Fig. 4. Conducting tests with a professional cricket bowler at BAC

To track the 3D acceleration of the bowler's arm as shown in Figure 5, the fully charged IMU was positioned as shown in Figure 1(b) and Figure 3.

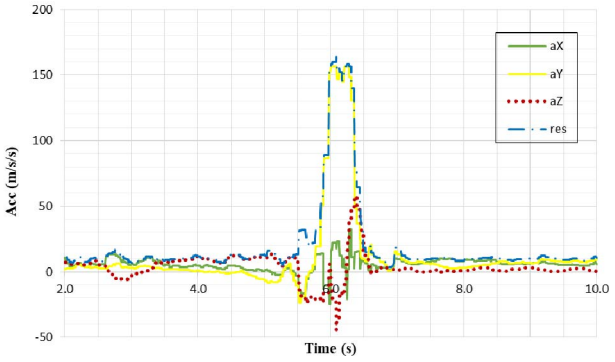


Fig. 5. The Recorded Acceleration data for the 7th delivery by the professional bowler.

An elastic sweatband was used to fix the IMU securely to the wrist of the bowler and this ensured alignment with the wrist axes. With such a position, the sensor-intrinsic y axis coincided with the longitudinal axis of the bowler's arm, while the x and z axes referred to the lateral and vertical arm axes respectively. The axes move with the arm during bowling. The run up was limited to 10 metres to maintain a Bluetooth connection between the IMU and the computer.

The data received from the IMU was processed by a Python script (program) and both the raw and processed data was logged into a file. The calculated speed was displayed in real time on the screen.

B. Data Processing Approach

The accelerometer in the IMU tracked the acceleration generated by the motion of the arm and temporarily stored those values. The sensor data was processed by a DMP and sent via Bluetooth to the computer as a continuous stream of 3D acceleration from the IMU. The resultant acceleration was calculated from these values of acceleration. The resultant was simply the square root of the sum of the squares of the acceleration in the 3 axes.

The basis of the derivation of speed from acceleration comes from elementary Physics. It is known that the rate of change of the velocity with respect to time is the acceleration of an object and hence the validity of Equation (1).

$$a = \frac{dv}{dt} \tag{1}$$

By fundamental definition, velocity is therefore the time integral of acceleration [14]:

$$V(t) = V_o + \int_{t_o}^{t} a(t)dt\,, \quad V_0 \text{ is the velocity at time } t_0 \tag{2}$$

Hence given an object's acceleration versus time graph, the integral of the graph gives the object's velocity at any given time interval. This concept is illustrated in Figure 6 and summed up by Equation 2. The integral of the acceleration within the given time interval is equal to the shaded area under the curve.

It has been established that the accuracy of IMU-based systems in velocity measurement rapidly degrades over time due to inherent sensor noises [15]. Therefore, the acceleration was integrated numerically over a minute period of time around the release point to get the speed using the Trapezoidal method of integrating data.

The algorithm for isolating the release time employed conditional statements. These statements together with the 'maxima' algorithms kept track of the value of acceleration in order to pinpoint the peak value of the acceleration upon deceleration of the acceleration values. Isolation of the interval over which to integrate involved comparing the resultant value to a threshold. This threshold was determined experimentally and set high enough to ensure that they are close to the peak acceleration for any given delivery. Calculations were only done when the resultant was greater than the threshold to eliminate the effects of running before the ball was delivered. The peak value of the resultant (also known as the global maxima) was found using the 'max' function of the 'numpy' module. Figure 7 displays a flow diagram detailing the steps followed in developing the Python software program.

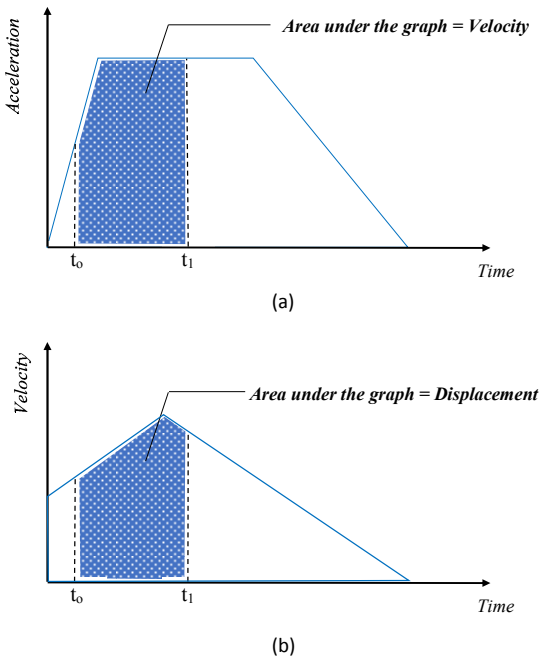


Fig. 6. Numerical Integration of (a) acceleration and (b) velocity

The algorithm for calculating the speed used a Numerical Integration technique [8]. The Trapezoidal rule was used to perform Numerical integration on the acceleration data. Trapezoidal integration was achieved using the 'trapz' method of the 'numpy' module.

Numerical Integration is used to find the integral in cases where an array (list in Python) of 'y' values is available but the functional dependence of 'x' and 'y' is unknown. The 'x' values are either specified or an increment 'dx' is given.

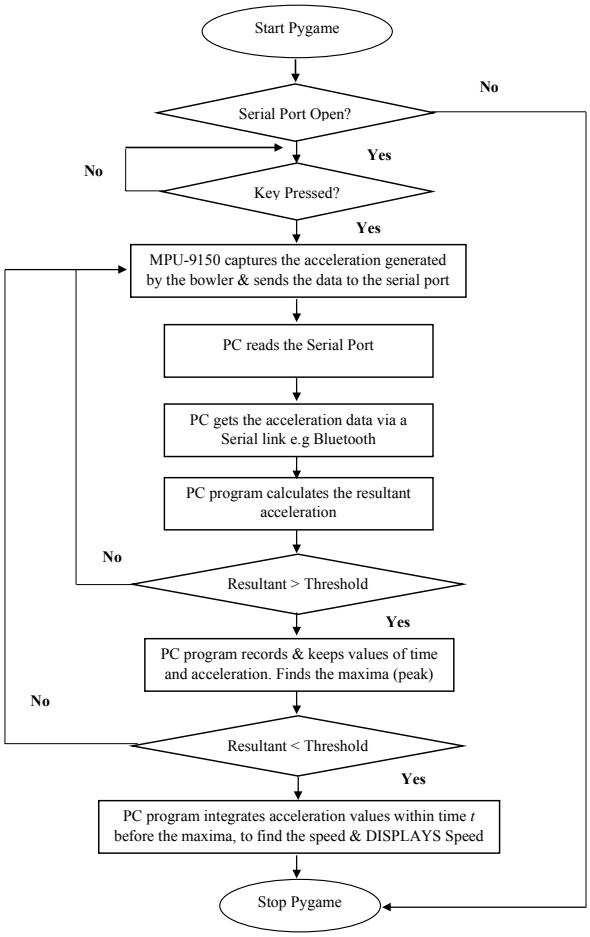


Fig. 7. Flowchart for developing the Python software

IV. Results and Discussion

The main output of this research was the determination of the speed at which the cricket ball was released. The WIBASE was able to fetch acceleration data from the IMU and calculate the speed. In addition to logging selected sensor data and calculated data onto a text file, the Python program displayed the speed on a Pygame window as shown in Figure 8. This display was updated in realtime by a method in Python known as 'blitting'. In which the text surface was continually rewritten on the display screen.

Fig. 8. Display of the calculates speed values in realtime

Table 1 shows the speed that was measured by the WIBASE for some balls delivered by the three participants during experimentation. Four experiments were chosen arbitrarily.

TABLE I. Results from deliveries by 3 different bowler

Experiment Number	Laboratory Speed (km/hr)	Amateur's Speed (km/hr)	Professional's Speed (km/hr)
1	91.804	102.527	116.561
6	85.65	120.104	139.304
10	104.858	108.291	123.349
12	105.38	109.269	132.18
Average Speed of 12 Experiments	***95.040167***	***110.8495***	***129.08808***

An informal interview was conducted with the professional cricket bowler to determine among other issues, the slowest ball he has recorded, the fastest ball he has delivered and measured and the average speed as determined during international matches. These known statistical values of the professional bowler were used as a reference to assess the validity of the system and the speed calculated as presented in Figures 9 and 10, following the failure to obtain permission to use the bowling speed measuring equipment in South Africa. The bowler revealed that the slowest deliveries are in the range 120-122 km/hr, meanwhile the fastest, recorded delivery is 145 km/hr and the average speed is between 130 km/hr and 135 km/hr., This method was good enough to pinpoint the precision of the system but only a standard system could determine the accuracy 100%.

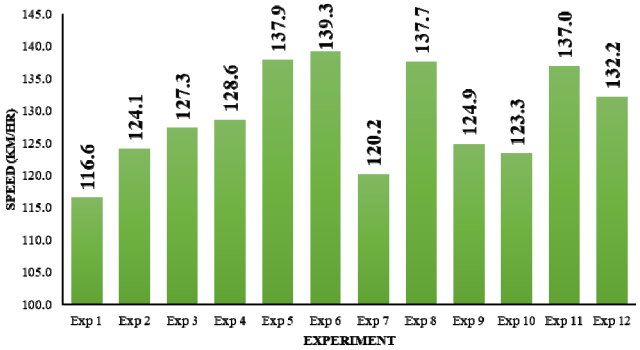


Fig. 9. The speed calculated for each of the 12 deliveries bowled by the professional bowler.

The crucial plots of the bowling speed measured by the system are shown in Figure 9. They serve to validate the system or disapprove it. The speed recorded when each of the 12 deliveries was bowled appear to increase gradually from the slowest during the first delivery to the fasted on the 6th delivery. This may be a consequence of the bowler enhancing their ability to bowl fast after warming up. Beyond the 6th delivery the speed fluctuates but remains relatively high. The slowest delivery was measured to be 116.56 km/hr while the maximum speed that was measured was 139.30 km/hr. The speed at release for the 2 overs bowled by the professional cricketer ranged from 116.6 km/hr to 139.3 km/hr. In some instances, the speed was quite slow by the bowler's standards because the intensity with

which one bowls in a game situation differs from a practice session.

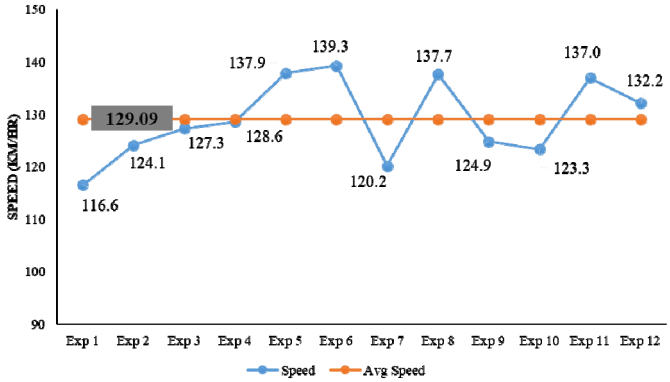


Fig. 10. Comparison of the actual speed and the average speed of the professional bowler

A comparison of the speeds of the 12 deliveries to the average speed of 129.08 km/hr calculated from the 12 values is presented in Figure 10. Save for the slowest delivery all the other values of speed fall within the 10% deviation bracket from the average speed. Some measure of consistency was realized from the performance by the professional cricket bowler.

The graphs in Figure 11 show a comparison between the bowling speed from 12 experiments conducted with a professional bowler and his documented statistics from international cricket matches.

The plots reveal that all the deliveries by the professional bowler neither reached nor surpassed the fastest delivery ever delivered by him. The match average speed was comparable to the average speed calculated from the 12 deliveries, it differed by a mere 3,91 km/hr. Two of the deliveries bowled during experimentation fell below the match slowest speed. All this can be attributed to the fact that the experiments were not conducted under the pressures, intensity or rigors of a match. The speeds bowled by the professional bowler were also compared to bowling speeds observed during cricket matches that were broadcast on television. The fast bowlers generally bowled in the range of 130 km/hr to 145 km/hr in real match situations. With this analysis in addition to that discussed earlier, the WIBASE was found to be a realistic solution to the measurement of bowling speed especially during training.

The system measures the speed relatively consistently. Since the speed of the 12 deliveries are dotted on either side of the average speed this further strengthens the argument for the consistency of the throws and probably the system. This trend makes sense because no individual bowler can bowl at the same speed over many overs however some consistency can be maintained more so over a duration of just two overs (12 balls).

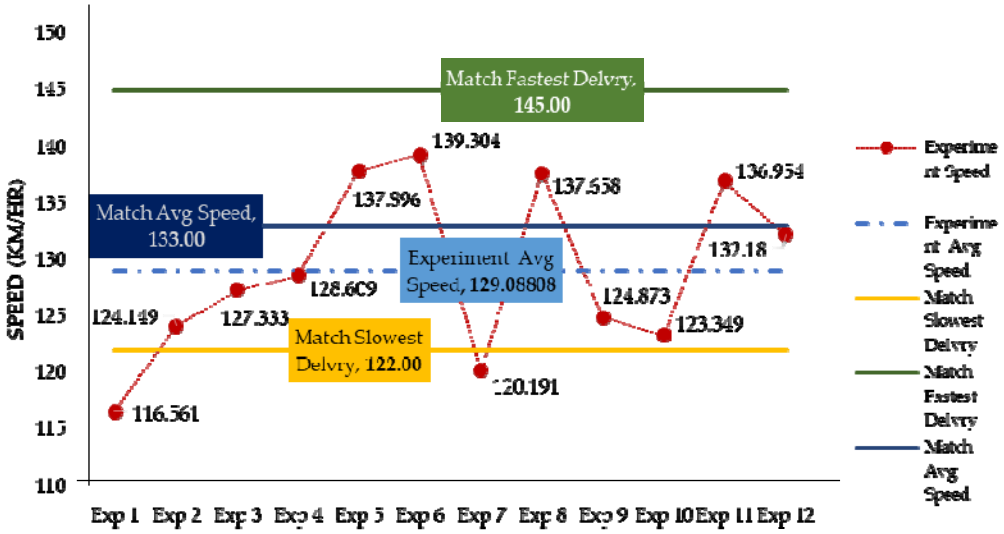


Fig. 11. Benchmarking the WIBASE using the bowling statistics of a professional bowler.

The classification of bowlers in accordance with their bowling speeds is presented in Table II. Fast bowlers bowl at speeds of 120 km/hr up to 161 km/hr. This table helps to further validate the results obtained from the experiments. The bowlers who took part in this research fell into these categories. The amateur bowler with an average speed of 111 km/hr was found to be a genuine medium bowling paceman and the professional bowler with his average speed of 129 km/hr was found to be a medium-fast paceman. The professional bowler could easily be classified as a fast-medium in a game situation were the pressure to perform increases the speed significantly.

TABLE II. Classification of speed based on the speed of delivery at release [15]

Bowler Classification	Ball velocity (m/hr)	Ball velocity (km/hr)
Fast	75-100	120 – 161
Medium	60-75	96.5 – 120
Slow (Spin)	40-60	64 – 96.5

To echo the sentiments of [3], the WIBASE can help coaches to pinpoint the strengths and weaknesses of athletes

during workout sessions and design an optimal personal training plan for athletes to improve their performance. The system can be used primarily to determine the speed of a cricket bowler and secondly it can be used to monitor the consistence of a bowler when bowling a long spell (many overs successively).

V. Conclusion

The feasibility of using MEMS accelerometers and a Python program to estimate the speed of a cricket bowler was demonstrated successfully. Consequently, a system that may be used by cricket bowlers and coaches to measure the bowling speed during training in order to enhance performance and consistency was developed. It was not possible to test the WIBASE against an existing and accepted method of determining speed, as alluded to in the results section but the approach chosen to benchmark the system indicates that the WIBASE may be a relatively good tool during training.

Acknowledgment

This research project was successfully undertaken because of resources acquired with the support of the Microlab in Switzerland and the National University of Science and Technology in Zimbabwe. The participation of Brian Chataika and Christopher Mpofu of Zimbabwe Cricket in the testing of the system and recording of measurements is appreciated.

References

[1] "ZC hunt for raw pacemen," Zimbabwe Independent, March 19, 2010. [Online]. Available: http://www.theindependent.co.zw/2010/03/19/zc-hunt-for-raw-pacemen/ [Accessed April 15, 2012].

[2] R. S. McGinnis and N. C. Perkins, "A Highly Miniaturized, Wireless Inertial Measurement Unit for Characterizing the Dynamics of Pitched Baseballs and Softballs," *Sensors*, vol. 12, pp. 11933-11945, August 2012.

[3] F. Dadashi, F. Crettenand, G. Millet and K. Aminian, "Front Crawl Instantaneous Velocity Estimation Using a Wearable Inertial Measurement Unit," *Sensors*, vol. 12, pp. 12927-12939, September 2012

[4] O. Mirabella, A. Raucea, F. Fisichella and L. Gentile, "A motion capture system for sport training and rehabilitation," *in Proc. of the 4th International Conference on Human System Interactions, HSI 2011, Yokohama, Japan, 19-21 May 2011*. pp. 52-59.

[5] S. B. Qaisar, S. Imtiaz, F. Faruq, A. Jamal, W. Iqbal, P. Glazier and S. Lee, "A Hidden Markov Model for Detection & Classification of Arm Action in Cricket Using Wearable Sensors," *J. Mobile Multimedia*, vol. 9, pp. 128-144, 2013.

[6] S. Yang, and Q. Li, "Inertial sensor-based methods in walking speed estimation: A systematic review," *Sensors*, vol. 12, pp. 6102–6116, May 2012

[7] H. E. Kevin and M. A. Darryl, "A Football wristband to measure throwing speed," Bachelors' thesis, University of Illinois, Chicago, USA, 2006.

[8] Slifka, L. D, "An Accelerometer based approach to measuring displacement of a vehicle body," Master of Science in Engineering, University of Michigan, Michigan, 2004.

[9] P. Kajánek, "Testing of the Possibilities of Using IMUs with Different Types of Movements," *Geoinformatics FCE CTU*, vol. 12, pp. 61-66, June 2014

[10] A. M. Sabatini, C. Martelloni, S. Scapellato and F. Cavallo, "Assessment of Walking Features from Foot Inertial Sensing," *IEEE Transactions on Biomedical Engineering BME*, vol. 52, pp. 486-494, March 2005

[11] W. Spratford, M. Portus, A. Wixted, R. Leadbetter and D. A. James, "Peak outward acceleration and ball release in cricket," *J Sports Sci,* vol. 33, pp. 754-760, April 2015

[12] *InvenSense* 2012, "MPU-9150 MotionFit Wireless SDK Application Note," Sunnyvale, USA.

[13] Q. Yuan, I. M. Chen and A. Caus, "Human velocity tracking and localization using 3 IMU sensors," in *Proc. Of the IEEE Conference on Robotics, Automation and Mechatronics, RAM, Manila, Philippines, November 12-15, 2013,* pp. 25-30.

[14] J. M. Cimbala, "Linear Velocity Measurement," Penn State University, 2013.

[15] D. H. Titterton and J. L. Weston, *Strapdown Inertial Navigation Technology*, Peter Peregrinus Ltd.: London, UK, 2004

[16] L. Justham, A. West, A. Harland and A. Cork, "Quantification of the cricket bowling delivery; a study of elite players to gauge variability and controllability." *The Engineering of Sport 6*, pp. 205-210, 2006.

Augmented-Reality Computer-Vision Assisted Disaggregated Energy Monitoring and IoT Control Platform

Jannish A. Purmaissur, Praveer Towakel, Shivanand P. Guness and Amar Seeam
Middlesex University (Mauritius)
Uniciti, Flic en Flac
Mauritius
Email: {j.purmaissur,p.towakel,s.guness,a.seeam} at mdx.ac.mu

Xavier A. Bellekens
University of Abertay
Dundee
Scotland
Email: {x.bellekens} at abertay.ac.uk

***Abstract*—The aim of this research is to develop an innovative low cost and affordable platform for smart home control and energy monitoring interfaced with augmented reality. This method will educate people about energy use at a time when fuel costs are rising and create novel methods of interaction for those with disabilities. In order to increase the awareness of energy consumption, we have developed an interactive system using Augmented Reality to show live energy usage of electrical components. This system allows the user to view his real time energy consumption and at the same time offers the possibility to interact with the device in Augmented Reality. The energy usage was captured and stored in a database which can be accessed for energy monitoring. We believe that the combinations of both, complex smart home applications and transparent interactive user interface will increase the awareness of energy consumption.**

***Index Terms*—Air Quality, Augmented Reality, Presence detection, Sensor**

I. INTRODUCTION

Future smart home platforms will be built upon various upcoming paradigm shifts in technology - namely Augmented-Reality, Computer Vision assisted detection, Energy Monitoring-Disaggregation and IoT Control (ACADEMIC). The ACADEMIC platform we propose will be based on the development of integration of these technologies in a low cost and affordable package. ACADEMIC is the first of its kind - albeit very similar propositions, or subsets of ACADEMIC have been proposed before. Our intention is to draw upon the latest research and create a smart home platform that is easy to use, and engaging for all generations.

Smart Homes have been around for quite some time, but have failed to capture the interest of people. The Internet of Things has however recently helped to introduce smart home concepts to people through cloud middleware platforms (e.g. IFTTT [1]), whereby simple control mechanisms are enabled through the internet (e.g. turning on a lamp at night automatically from the internet or via a motion sensor, yielding potential savings of up 20% [2]) though complete systems of whole house automation and control are still treated as esoteric or of domain of the wealthy [3]. Smart Homes though have enormous potential to make our lives better - they can enable us to have better control over our homes (particularly at the assistive technologies level [4]), whilst provide us multiple monitoring data which can be analysed further [5](e.g. how much does my air conditioner cost to run?, will it be cheaper to do 'x' versus 'y'?). Moreover home owners have a high interest in reducing their energy consumption as it can be considered an important cost factor. However standard electricity meters and analogue billing systems lack the feedback to create energy awareness. This is where smart home control and energy monitoring interfaced with augmented reality will help to shorten the feedback time from consumption of energy to user billing system and at the same time increase users energy consumption awareness.

II. COMPUTER VISION

Computer vision has been used in Smart Homes for monitoring purposes and offer ways to interact its appliances. Different ways to interact with the house include using a device such as a smart phone [6], [7], [8], eye gaze [9] and hand gestures [10], [11], [12], [13]. Smart home should also be able to monitor the activity [14], [15] and the well being [16], [17] of the inhabitants. Computer vision enables the smart home to be more accessible to users with different capabilities, it enables the inhabitants to have an easy way to interact with their surrounding and contributes to improvement of their quality of life [18].

Jafri et al.[15] developed an assistive device using Google Tango tablet to enable people who are visually impaired to navigate indoors and avoid obstacles. The depth camera in Tango is used to detect objects in the path of the user and a speech based navigation information is given to the user.Takizawa it al.[19], proposed a system to help people who are visually impaired to search and find objects using Microsoft Kinect. The system was used to measure the depth of field to find objects such as chairs, floor, downwards and upwards stairs. In the experiment carried out, it was found that the time to find the object was significantly shorter than that of a conventional white cane.

III. METHODOLOGY

A. Node-Red and Augmented reality

The system was built on top of the Node-Red[20] framework running on a Raspberry Pi. Node-Red is a flow based development tool developed for wiring hardware devices, APIs (Application programming interface) and on-line services as part of Internet of Things(IoT). Moreover different communication protocols can be used for interconnection of heterogeneous devices for example an Arduino connecting to a Raspberry Pi. Node-Red was integrated into the smart plug network in order to provide real time data. Node-Red provides developers the facility to build scalable embedded systems and at the same time provides a web service to control different devices irrespective to its network technology. Node-Red comprises of different node which can be interconnected to communicate with different devices and protocols.

B. Influx DB

Influx DB[21] is an open source database optimized for fast, high-availability storage and retrieval of time series data in fields such as operations monitoring, application metrics, IoT sensor data, and real-time analytics[22].The data recorded in Influx DB are displayed in Grafana[23].

C. TP-Link HS110 integration

Energy consumption monitoring devices should provide a maximum transparency. TP-Link HS110[23] wireless smart plug was used to receive energy consumption data of plugged-in devices. TP-link HS110 connects to the wireless network and provides information such as current, power, voltage and many more. To integrate the TP-Link HS110 functionality, Node-Red flows were created inside the Node-Red environment.

D. OpenCV

OpenCV[24] is an open source computer vision and machine learning software library. This library contains over 2500 algorithms.

E. Python

Python[25] is an interpreted high-level programming language for general-purpose programming.

F. Raspberry Pi Camera

The Raspberry Pi camera we used was an 8 mega-pixel Sony IMX219. This is widely available and low cost.

G. Integrated System Architecture

Figure 1 provides a flow diagram of the system. The architecture flow is comprised of 3 main components: the TP-Link 110, Raspberry Pi and Augmented Reality application. These components are all interdependent. The Raspberry Pi is a micro-controller hosting the basic infrastructure like Influx DB, Node-Red etc.The proposed application aims at integrating energy consumption into a smart home infrastructure and providing user interface to interact with the smart home system. A mobile interface allows the user to control the smart plug and visualize the live data. Energy consumption is monitored using a TP-LINK HS110 smart plug connected to the smart home wireless network. On their mobile devices, users can directly access the smart plug via a node-red dashboard. It recognizes the smart plug using image processing techniques and displays energy consumption of the device. The data captured from the smart plug are stored in an Influx DB database which is then requested by a Grafana interface. Furthermore, the application enables the user to view the data on both a Node-Red dashboard and the Grafana interface. In future, smart homes energy efficiency will be a major issue, when considering advancements in the area of smart metering and smart grids [26]. For example, time-of-use and real-time pricing may help both, consumers to save energy costs in households and energy providers to improve load management[27]. Moreover, it would be possible for the user to know which type of energy he is using such as renewable or nonrenewable sources. For example the user can be rewarded when using green energy and thus decreasing his monthly bill. Thus, energy aware smart home applications has to find a balance between supporting the user in saving energy and at the same time not decreasing convenience[6]. A real time monitoring system is provided with the integration of a database system to keep a detailed record of the energy usage.

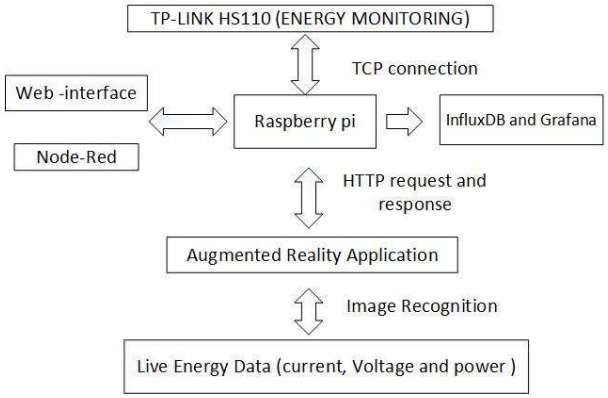


Fig. 1. Architecture flow

H. Augmented Reality and Computer vision in Air Quality Monitoring

For the computer vision part of the project we went with a Raspberry Pi camera which is both reliable and easy to use.This was used in conjunction with OpenCV to capture and analyse a stream of pictures to detect motion in the room. This allow the us to determine motion in the room.

I. Occupancy tracking

In this section, we detail how we were able to determine the occupancy of a room using OpenCV and a Raspberry Pi camera.

1) Algorithm: OpenCv and Python were used to detect motion in a room. We also used a JSON(JavaScript Object Notation) file instead of line arguments. This allowed for quick and easy set-up for the python scripts. The classes we used from the PyImageSearch package are as follows.

- TempImage:This allow for temporary images to be stored in a buffer
- PiRGBArray:Allow for array manipulation of RGB
- PiCamera:Allow for capture of a stream of images from the camera

To start image collection, we set-up our Raspberry Pi Camera and allow for a warm up time of up to a few seconds to ensure that the capture-sensors are given enough time to calibrate. We also initialize the average background frame. The first process is to re-size the frame to a width of 500-pixels. This is done to have standard video output. Following this process we convert the images to gray-scale and apply a Gaussian blur to remove high frequency noise which allow us to focus on the "structural" objects of the image. This is shown in 2.

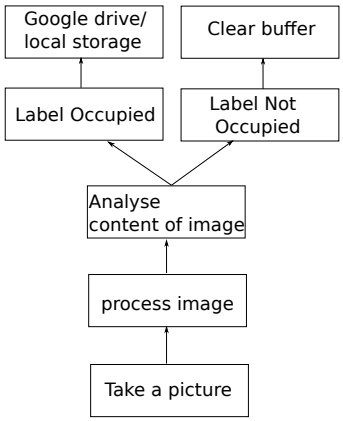


Fig. 2. Motion detection sequence

2) Weighted average: The first frame of the stream is not always a good representation of the background we want to model, factors such as lighting condition change or people entering room can affect the model. For this reason we instead take a weighted mean of the previous frame along with the current frame. This means the python script dynamically adjust to the background at different times in the day. Thus creating a better model for the background in relation to the foreground. Based on the weighted average of frames, we then subtract the weighted average from the current frame. This result in

$$delta = |background\_model - current\_frame| \quad (1)$$

3) Threshold: We can then threshold this delta to find region of the image that contains difference from the background. These areas are then detected as motion in the video stream.

4) Contour detection: The areas in the image that pass the threshold test are then subjected to contour detection. If the region are big enough then we indicate that we have found motion in our current frame. We then compute a bounding box of the contour and draw a box around the motion.

J. AR in air quality monitoring

AR can have many practical uses. One of them is displaying data. In the smart home there are a multitude of sensors typical example are temperature and humidity. Recently air quality has been a variable of interest as we seek healthier lifestyles and also have been added to smart homes systems. Our current AR system has been developed to interact with bespoke air quality monitor systems. Structure of the node is shown in 3

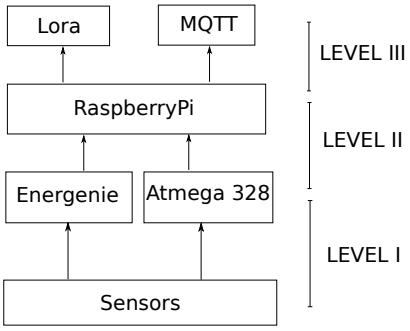


Fig. 3. Air-quality node flow-diagram

This system would work like this

1) One time Download of the application
2) Point at the air quality system
3) Read data for the system

Fig. 4. Air quality in Augmented Reality display

The use of AR, as shown in figure 4, make the whole thing very intuitive to the user. Not only does it allow for display of data, it also shows a breakdown of the different sensor data. This does not stop at regular user though. Performing maintenance on these devices also becomes easier. For someone doing maintenance of these devices, a facilities manager for example could go and check the status of the building sensor with only their smart phone, this makes it easier to work in the fields. This allows for the quick debugging and the spotting of faulty sensor nodes. Another aspect of the AR here is the ability of a superuser(maintenance officer) to change the system software. This would require no cable connection or cumbersome VPN or VNC connection to the device, whereby the officer could point the smart phone at the device to be updated.This could mean quicker, more frequent updates to the devices in a building.

IV. Results

Currently the energy consumption values are saved into a database and the user has to access it manually in order to visualize the data. This becomes quite complicated and cumbersome for the user to access the database each time he wants to visualize. Therefore, it is important to give the user the ability to visualize his data in an easy and interactive manner. In the following, we present interaction techniques which the user can uses to visualize his energy consumption information.

A. Results of OpenCV

In this section, we show results from our python script detecting motion in a room. First a stream of images is taken from the Pi-Camera. Then this stream of image was processed to get the gray-scale and the Gaussian-blur. The result is shown in figure 5.

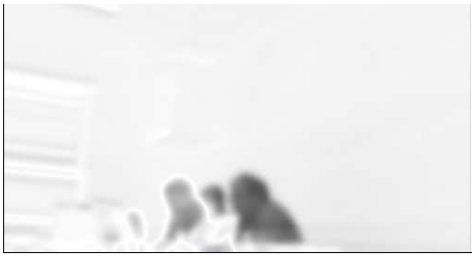


Fig. 5. Inverted reference frame for motion detection(Gaussian blur)

This image can now be used as a reference frame for future captures. As new frames are captured they are compared to the reference image(Gaussian-blur) by thresholding. Contouring is done to box the motion areas in the frame and we end up with figure 6. The tracking system proved to be reliable in detecting motion and in labelling the room as occupied when actually occupied.

Fig. 6. Motion detection

B. Application

A direct application of the motion detection could be the switching of ceiling lights of segments in a room by mapping lights position. In its final implementation this could conserve energy as it would only light relevant segments in the room.

Fig. 7. Augmented reality data

C. Augmented Reality

The easiest way of receiving information from a device in the real world is directly from the device. For this to be possible, the device needs to be augmented with information. Thus an intermediate device such as a smart-phone needs to be used to make the information virtually visible for the user. For this purpose, we have used a smart-phone to allow the user to see augmented data on their screen. Figure 7 and Figure 8, shows a working prototype where data is projected into Augmented Reality and Figure 9 shows data captured from the smart plug into Grafana dashboard when the database button is pressed. MQTT is used to send the data into the format of a JSON file that include current, power voltage, gas, $PM_{2.5}$, PM_{10}, temperature, pressure and humidity readings from air quality monitoring units.

Fig. 8. switched ON

The concept demonstrated above is using a smart-phone running android[28] and Vuforia[29] with an internal image recognition database.The camera image of the mobile phone is compared to the image recognition database. It recognizes the image by using trackers. Finally, the mobile phone request the live data from Node-Red server along with its GPS location, the return data is based on the specific coordinates provided.

Moreover, the user is also able to interact with the object through the mobile phone. When the object is switched on, an HTTP response is sent to the Node-Red server which switches the smart plug on and increases the power of the smart plug. Figure 8, shows an example of the plug switched on. The same process is used to turn the power off.

The advantage of Vuforia interaction technique is its intuitive and easy handling. It uses Computer Vision to recognize and track planar images .Its image registration capability allows it to track the position and orientation of the image in real time.

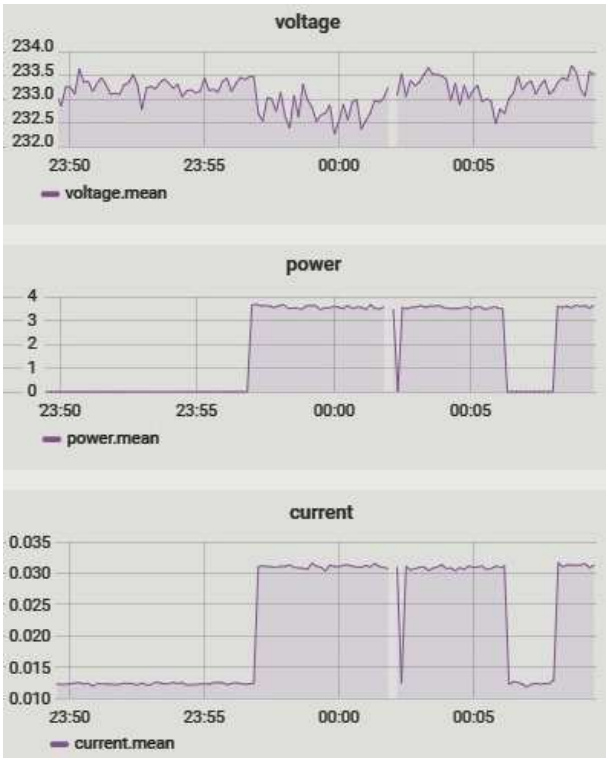


Fig. 9. Grafana dashboard

V. CONCLUSION

This project represent the beginning of our research work on the ACADEMIC platform. We discussed in this paper some of the technologies that were present in our system.

During the past few years, the adoption of electric and electronic devices has brought immense benefits to human beings, such as increased productivity, improved entertainment and communication and much more.

The increase of energy consumption awareness in the society will lead people to utilize energy more efficiently and consciously. We have brought this concept closer to the goal of increasing energy wastage awareness. The users are able to visualize their energy consumption and based on this knowledge they are in a better position to implement strategies for energy saving.

As smart home systems becomes smarter, they will be able to develop energy saving strategies for the user such as selectively switching of the lights.

The challenges discussed in this paper show that there are many more opportunities for further research. For the current study only one smart plug and air quality sensors are being used. For further development, we plan to integrate several smart home devices into Augmented Reality and work with Computer Vision instead of image recognition. Moreover we are also planning to optimize the performance of energy consumption using disaggregated energy monitoring system and concentrate on the overhead of the monitoring system in terms of energy usage/costs. The prototypes will be deployed in a classroom/lecture theatre environment both for monitoring air quality and also to provide students with information about indoor/outdoor air quality being measuredg with our existing sensor nodes.

REFERENCES

[1] B. Ur, M. P. Y. Ho, S. Brawner, J. Lee, S. Mennicken, N. Picard, D. Schulze, and M. L. Littman, "Trigger-action programming in the wild: An analysis of 200,000 ifttt recipes," 2012.

[2] R. Woodbury, L. Bartram, R. Cole, R. Hyde, D. Macleod, D. Marques, T. Mueller, and D. Vanier, "Buildings and climate solutions," *Pacific Institute for Climate Solutions*, 2008.

[3] A. Seeam, J. Liao, L. Stankovic, and V. Stankovic, "Improving energy efficiency with smart home appliance decision support tools," in *7th International Conference on Energy Efficiency in Domestic Appliances and Lighting (EEDAL)*, 9 2013.

[4] M. E. Morris, B. Adair, K. Miller, E. Ozanne, R. Hansen, A. J. Pearce, N. Santamaria, L. Viega, M. Long, and C. M. Said, "Smart-home technologies to assist older people to live well at home," *Journal of aging science*, vol. 1, no. 1, pp. 1–9, 2013.

[5] S. K. Firth, F. Fouchal, T. Kane, V. Dimitriou, and T. M. Hassan, "Decision support systems for domestic retrofit provision using smart home data streams." CIB, 2013.

[6] M. Jahn, M. Jentsch, C. R. Prause, F. Pramudianto, A. Al-Akkad, and R. Reiners, "The Energy Aware Smart Home," in *2010 5th International Conference on Future Information Technology*, May 2010, pp. 1–8.

[7] Y. Kim and Y. Lee, "[ieee 2015 ieee 29th international conference on advanced information networking and applications workshops (waina) - gwangiu, south korea (2015.3.24-2015.3.27)] 2015 ieee 29th international conference on advanced information networking and applications workshops - automatic generation of social relationships between internet of things in smart home using sdn-based home cloud," 2015. [Online]. Available: http://gen.lib.rus.ec/scimag/index.php?s=10.1109/WAINA.2015.93

[8] L. Z. W. Tang, K. S. Ang, M. Amirul, M. B. M. Yusoff, C. K. Tng, M. D. B. M. Alyas, J. G. Lim, P. K. Kyaw, and F. Folianto, "Augmented reality control home (ARCH) for disabled and elderlies," in *Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2015 IEEE Tenth International Conference on*, Apr. 2015, pp. 1–2.

[9] D. Bonino, E. Castellina, F. Corno, L. D. Russis, and L. De Russis, "DOGeye: Controlling your home with eye interaction," *Interact. Comput.*, vol. 23, no. 5, pp. 484–498, Sep. 2011.

[10] M. Budde, M. Berning, C. Baumgärtner, F. Kinn, T. Kopf, S. Ochs, F. Reiche, T. Riedel, and M. Beigl, "Point & control – interaction in smart environments: you only click twice," in *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*. ACM, 8 Sep. 2013, pp. 303–306.

[11] D.-L. Dinh, J. T. Kim, and T.-S. Kim, "Hand gesture recognition and interface via a depth imaging sensor for smart home appliances," *Energy Procedia*, vol. 62, pp. 576–582, 1 Jan. 2014.

[12] K. S. C. Kumar, "Stereo-vision based smart TV control," in *2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS)*, Nov. 2015, pp. 67–71.

[13] A. Mohammed, A. Seeam, X. Bellekens, K. Nieradzinska, and V. Ramsurrun, "Gesture based iot light control for smart clothing," in *Emerging Technologies and Innovative Business Practices for the Transformation of Societies (EmergiTech), IEEE International Conference on*. IEEE, 2016, pp. 139–142.

[14] J. Han and J. Han, "RGB-D human identification and tracking in a smart environment," in *Computer Vision and Machine Learning with RGB-D Sensors*, ser. Advances in Computer Vision and Pattern Recognition, L. Shao, J. Han, P. Kohli, and Z. Zhang, Eds. Springer International Publishing, 2014, pp. 195–211.

[15] R. Jafri and M. M. Khan, "Obstacle detection and avoidance for the visually impaired in indoors environments using googles project tango device," in *Computers Helping People with Special Needs*, ser. Lecture Notes in Computer Science, K. Miesenberger, C. Bühler, and P. Penaz, Eds. Springer International Publishing, 13 Jul. 2016, pp. 179–185.

[16] M. Pediaditis, M. Tsiknakis, V. Kritsotakis, M. Gralczyk, S. Voutoufianakis, and P. Vorgia, "Exploiting advanced video analysis technologies for a smart home monitoring platform for epileptic patients: Technological and legal preconditions," in *Telecommunications and Multimedia (TEMU), 2012 International Conference on*, Jul. 2012.

[17] L. Y. Mano, B. S. Faial, L. H. Nakamura, P. H. Gomes, G. L. Libralon, R. I. Meneguete, G. P. Filho, G. T. Giancristofaro, G. Pessin, B. Krishnamachari, and J. Ueyama, "Exploiting IoT technologies for enhancing Health Smart Homes through patient identification and emotion recognition," *Internet of Things\: Research challenges and Solutions*, vol. 8990, pp. 178–190, Sep. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0140366416300688
[18] A. Fernandez-Caballero, "A grand challenge for vision systems: improving the quality of life and care of aging adults," *Frontiers in Robotics and {AI}*, vol. 2, 2015.
[19] H. Takizawa, S. Yamaguchi, M. Aoyagi, N. Ezaki, and S. Mizuno, "Kinect cane: an assistive system for the visually impaired based on the concept of object recognition aid," *Pers. Ubiquit. Comput.*, vol. 19, no. 5-6, pp. 955–965, 26 Feb. 2015.
[20] D. Guinard and V. Trifa, *Building the web of things: with examples in node. js and raspberry pi*. Manning Publications Co., 2016.
[21] A. Bader, "Comparison of time series databases," Master's thesis, 2016.
[22] D. Namiot, "Time series databases." in *DAMDID/RCDL*, 2015, pp. 132–137.
[23] D. Thalmann, "An interactive data visualization system," *Software: Practice and Experience*, vol. 14, no. 3, pp. 277–290, 1984.
[24] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*. " O'Reilly Media, Inc.", 2008.
[25] G. Rossum, "Python reference manual," 1995.
[26] M. Webb *et al.*, "Smart 2020: Enabling the low carbon economy in the information age," *The Climate Group. London*, vol. 1, no. 1, pp. 1–1, 2008.
[27] S. Darby, "The effectiveness of feedback on energy consumption," 2006.
[28] A. Developers, "What is android," 2011.
[29] A. K. Wahyudi, R. Ferdiana, and R. Hartanto, "Pengujian dan evaluasi buku interaktif augmented reality arca 3d," *SEMNASTEKNOMEDIA ONLINE*, vol. 2, no. 1, pp. 2–07, 2014.

Prohibition Signage Classification for the Visually Impaired Using AlexNet Transfer Learning Approach

Kefentse Motshoane
Department of Computer Systems Engineering
Tshwane University of Technology
Pretoria, South Africa
kzm.motshoane@gmail.com

Chunling Tu
Department of Computer Systems Engineering
Tshwane University of Technology
Pretoria, South Africa
tclchunling@gmail.com

Pius Adewale Owolawi
Department of Computer Systems Engineering
Tshwane University of Technology
Pretoria, South Africa
owolawipa@tut.ac.za

***Abstract*—Prohibition signs are commonly used for safety purposes in order to prevent and protect individuals from dangerous situations. These signs are placed in or around areas whereby they are clearly visible to the public. However, the visually impaired cannot visualize such signs. To help them, this paper proposes a system that combines Convolutional Neural Network (CNN) model and Computer Vision (CV) algorithms to detect and recognize prohibition signs in real scenes. The system uses pre-trained AlexNet model, fine-tuned using Prohibition Signage Boards (PSB) dataset and combined with Maximally Stable Extremal Regions (MSER) and Optical Character Recognition (OCR) techniques for text extraction and classification, to enhance the system performance. The experiments indicate that high recognition accuracies are achieved from a variety of prohibition images and prohibition texts.**

Keywords—CNN, AlexNet, Transfer Learning, Computer Vision, MSER, OCR

I. INTRODUCTION

The visually impaired community come across numerous challenges when performing their daily activities. They are highly dependent on the assistance from sighted individuals in visualizing objects around them. Dependency might be a disadvantage, as the visually impaired people must independently account to prohibition signage boards (PSB) in order to prevent and protect themselves from dangerous situations. The use of cameras has been dominant as the primary solution in vision substitution (VS). Cameras alone are limited to capturing visuals of the scene, but fail to process, analyze and classify features within the scene [1].

Several research studies have produced state-of-the-art solutions to various research problems under the domain of VS using cameras. Over the years, CV algorithms such as Scale Invariant Feature Transform (SIFT), Speed Up Robust Feature (SURF) and Maximally Stable Extremal Regions (MSER) were commonly used in developing artifacts to aid the visually impaired to detect features and extract text information in their surrounding environment [2]. Many image classification methods for VS systems were developed based on Convolutional Neural Networks (CNNs). In this paper we propose an approach that utilize a pre-trained CNN using AlexNet architecture to classify PSB. AlexNet is a large CNN which was reported to successfully classify 1.2 million high-resolution images using 1000 different classes [3]. The AlexNet model is also an efficient deep neural network model designed for object detection and classification [4]. Recent and current state-of-the-art CNNs includes GoogLeNet [5] and RestNet, winners of ILSVRC 2014 and 2015 respectively. ResNet, which was developed and trained on a subset of the ImageNet database, has demonstrated higher performance in image classification [6].

In this paper, to accommodate the specific purpose to aid visual impaired, the proposed approach eliminates the pre-trained classes that were shipped with the CNN model. The model is re-trained using four different classes of our prohibition sign board dataset, which includes 500 images per class. During the training process, the performance of the model is monitored, recorded and analyzed based on the classification accuracy. The limitation of the CNN models is addressed. The models focus on image classification; however, for the specific purposes of our research, the text from sign boards should be extracted in order to provide more detailed information to the VI. To address this limitation, the proposed system combines CNN capabilities of classification with CV MSER and OCR techniques, which have been proven to be capable of classifying sign boards, and detecting and extracting text on the boards.

II. RELATED WORKS

The PSB are commonly based on symbols as form of visual communication but they also display text information. Algorithms were developed to classify symbols based on image and text information.

A. Symbol detection and classification

CNNs have been known for their significant performance in applications based on visual tasks processing and natural language processing [7]. They inherited their name from the convolutional layers that form part of their architectural blue print [3]. CNNs can be used as a highly-accurate method for image classification. These models are complex but efficient with a high rate of discrimination and have proven to provide good results in image classification, object detection, and fine-grained classification. CNNs are distinguished based on their architecture and each architecture can produce results that vary from others depending on a specific function or study they are subjected to execute. It is a common practice for CNNs to maintain an architectural design which includes a set of stacked layers executing both linear and non-linear process. [8]. A study to investigate three classes of skin lesion classification problem using a low quality and small

size dataset was conducted using AlexNet transfer leaning approach. AlexNet archived an accuracy of 98.67% when a larger dataset size was used [7]. Another study was conducted for a tennis collection robot, which utilized AlexNet model for tennis recognition. The model was fine-tuned on tennis representation images, which were grouped in namely training sets, test sets and verification sets made up of tennis and non-tennis images respectively. The results have demonstrated that after the iteration of 10 000 operations, the accuracy achieved was 99% [9]. In a later study, Speech Emotion Recognition task was formulated as an image classification problem. The task is to automate the process of recognizing emotional aspects of speech irrespective of actual sematic contents using a computer software. The first part of the experiment tested the efficiency of AlexNet-SVM (Support Vector Machine) approach using RGB images of speech spectrograms as input to the pre-trained model. The second part tested the efficiency of Fine Tuned AlexNet (FNAlexNet) model using RGB images of speech spectrograms as input. General observations indicate that the average accuracy for FTAlexNet was about 79% and for AlexNet-SVM was about 68% [10].

B. Text extraction and classification

Many techniques were to extract and classify text, such as edge detection, region detection, region extraction and template matching. MSER algorithm was used for text region detection and extraction. OCR was used to input the extracted text regions with the aim to recognize and classify text by identifying patterns corresponding to alphanumeric or other characters [11]. The effective combination MSER and OCR techniques for text detection and extraction is further proven by [12]. The paper proposed an accurate and effective algorithm for detecting and recognizing any text in a natural image. The proposed algorithm required enhancing MSER by combining it with canny edge algorithm and applying strike width to eliminate the regions that contains too much variations. The remaining regions were taken as input to OCR to make the text usable. The evaluate results were 77.47% of the f-measure on the ICDAR 2011 dataset [12].

III. PROPOSED METHOD

In this section, the proposed method is introduced. The system flow diagram is illustrated in Fig. 1. The system is broken down into two main approaches: PSB symbol classification and text classification. The proposed method utilizes pre-processed PSB dataset. AlexNet CNN model is used in this research, for its architecture and the ability to transfer knowledge between different applications using transfer leaning approach. The text extraction and classification methods are presented, following the discussion on the limitation of CNNs for textual content classification. A small PSB dataset of real life scenes is used as input to evaluate the effectiveness of the system.

A. System Architecture

PSB dataset is used as input for the fine-turning process. The dataset is split into 70% training set and 30% test set. Each set is preprocessed to comply with the model input requirements. AlexNet model is the system center of our proposed approach, which executes two phases.

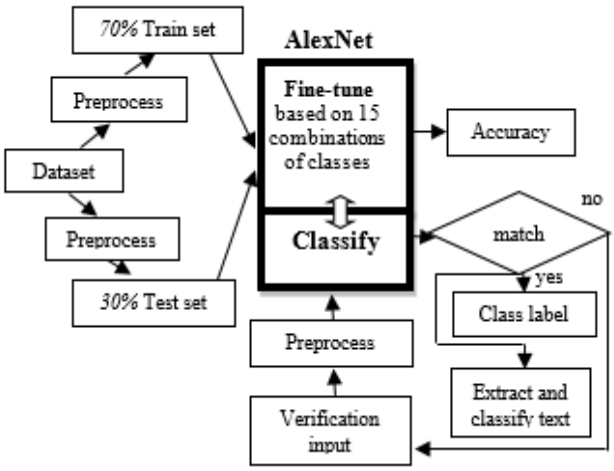


Fig. 1. System flow diagram

The first phase, the model receives the two sets of data, and executes the transfer learning process. The transfer learning processes requires only the last pre-trained fully connected layer to be replaced with new PSB classes.The model will use the weights (numerical values within an array representing a region of an image) obtained from the training set to compare and match with the weights obtained from the test set. Once the comparison process between the training and tests is complete, depending on the number of successful matches an accuracy will be computed. The testing process will be executed per combination of classes, and produce an accuracy per combination. The second phase of the system validates the system functionality. Sample images of PSB verification dataset representing the scene are provided as verification input to the model. The model executes a comparison process between the weights that the model was training on and the weights obtained from the verification input. Due to four number of classes, softmax function within the model is used to classify the verification input to a class with the most number of matches. If a classification is not made, the process restarts. Once a classification is made, the class label is provided as output and then text extraction process executes. Text is extracted using MSER and text classified using OCR. In the classification processed text extracted is verified by means of checking whether a class label which can also be referred as 'key words' exist in the text extracted. If key words are found, the process can be declared as successful.

B. Data Description

1) PSB category and design

PSB design varies in terms of symbols and text represented on the boards. Fig. 2, illustrates samples of images that represents PSB which fall under the four categories that were selected for this study, which includes 'No Pets', 'Do Not Drink', 'No Smoking' and 'No Entry' respectively.

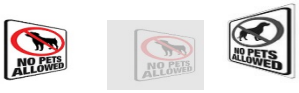
No Pets Do Not Drink No smoking No entry

Fig. 2. Standard Prohibition Signs (Source: Google).

2) PSB Augmentation

In most instances humans have the ability to identify and classify an object even when it is placed at different angles, resized, obscured, modified, etc. The ability comes naturally for humans, but when it comes to a computer based model, it is a different analogy. Computers are built in such a way that, if two types of data representing a single object, with one slightly modified, the computer would classify the data as distinct. In order to train a computer to classify multiple representation of the same object as indistinct, it is necessary to develop a training set that contain multiple varying representation of a single object, as illustrated on Table I.

TABLE I. MULTIPLE REPRESENTATION OF A SIGN BOARD.

Class	Original	Augmented samples
1		
2		
3		
4		

3) PSB Verification

PSB verification dataset are real images of a scene whereby a prohibition sign board may or may not exist, as illustrated on Fig.4. They are used to verify that the actual system is able to produces the expected results, in terms of detecting and classifying that the symbol exists in the scene and that the text written on the board can be extracted, classified and verified.

Fig. 4. Sample PSB verification dataset (Source: Google)

C. AlexNet and Transfer Learning Approach

AlexNet is a CNN which may be considered as a highly-accurate method for image classification. A typical CNN structure can be broken down to three parts: input layer, hidden layers and output layers [13], as illustrated on Fig. 5.

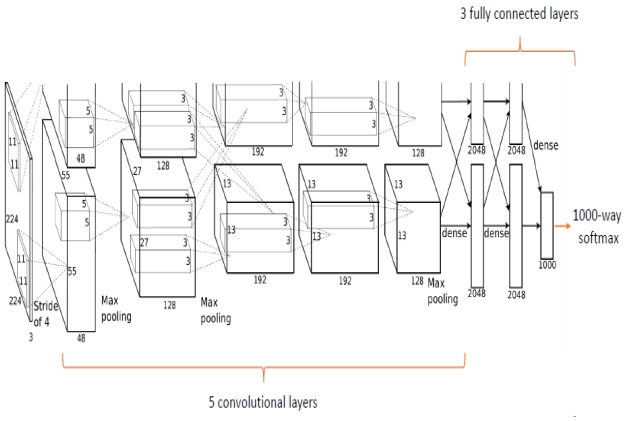


Fig. 5. AlexNet Architecture [4]

PSB datasets have to undergo the preprocessing phase in order to comply with the AlexNet input data requirements of 227x227x3 dimensions. The structure consists of convolutional layers to detect the local connection characteristics of the previous layer, pooling layers to fuses semantic similar features. Fine-tuning begins with copying (transferring) the weights from a pre-trained AlexNet model to a modified AlexNet model trained on our PSB dataset.

During the training process the aim is to replace the last fully connected layer of the pre-trained CNN with a new fully connected layer and the last fully connected layer will have the number of nodes equal to the number of classes in our PSB dataset [3]. For the purpose of generating the desired output we use softmax layer, which is a function within the CNN that maps output feature maps from the last fully connecter layer with a vector of real values (at times referred as weights) which represents the probabilities of each class [10].

D. MSER and OCR Text Detection and Classification

We use the fine-tuned AlexNet model, with sample images from PSB verification dataset representing a scene as input, in order to verify system functionality. The model obtains new sets of weights from the input image. It then compares and match them with the weights it was trained on. Once a sufficient number of matches is found, then the model will use the softmax function to label the input data to a class which resulted in the most number of matches. If a match is found, the system evokes a parallel MSER algorithm which performs feature detection and extracts of a number of co-variant regions that represents connected component of a pixel to the outer boundary pixel from the verification image. The extracted regions are being operated on the selected text region to recognize the actual text information of the image by using OCR algorithm [12]. The system results in two outcomes, the class label which the verification input belongs to and the text written on the verification input, based on successful classification and availability of textual content.

IV. EXPERIMENT AND RESULTS

In this section, we implement proposed method. The focus is firstly discussing the environment used for the experiment. Then followed by the setup and settings for the execution of the transfer learning approach. Results are recorded based on the accuracy obtained using combinations of classes. Once fine-tuning is complete, we then input verification data to test the classification and text extraction

process. The section is concluded by a brief discussion of the observation.

A. *Experiment Setup*

MATLAB 2018a was the preferred testing environment due to its comprehensive and available toolboxes in neural networks and computer vision. A total of 2000 images were used and 500 images per class. These images are categorized under the class labels.. Since there is a lack of large-scale PSB datasets available, some of free online images are collected individually from the internet. The dataset contains 80% of images, which were generated by augmenting the original images to produce multiple viewpoints and textural variety of the image.

B. *Fine-Tuning AlexNet Model*

During the implementation of our proposed method, we only modify the last pre-trained fully connected layer, which has 1000 classes and replace them with our 4 PSB classes. We train the network using 70% training set (TrS) and 30% test set (TsS) across all classes. During training the Stochastic Gradient Decent with Momentum (SGDM) method was used [14], the learning rate was set to 0.01, max epochs was set to 20 and mini batch size was set to 64.

The model is trained based on multiple combinations (cb) of PSB classes. The total number of combination that can be generated is 15, with the last combination consisting of all four classes producing an accuracy of 95% compared to 100% when the model is trained using a single class. Each combination produces an accuracy that is recorded as illustrated on Table II. Based on the results, it is visible that there is a drop in accuracy that occurs based on how the classes are combined during training. It is clear that the number of classes used to train the model does not impact much of the accuracy compared to the combination of classes. From the results, 94% is the lowest accuracy that was recorded and it was produced by a combination of two classes.

TABLE II. ACCURACY RESULTS FOR FOUR CLASSES OF PSB

	DATASETS		CLASSES				MODEL
c b	*No. TrS*	*No. TsS*					*Accuracy*
1	350	150	0	0	0	1	100%
1	350	150	0	0	1	0	100%
2	700	300	0	0	1	1	95.5%
1	350	150	0	1	0	0	100%
2	700	300	0	1	0	1	94%
2	700	300	0	1	1	0	96%
3	1050	450	0	1	1	1	95.3%
1	350	150	1	0	0	0	100%
2	700	300	1	0	0	1	99.5%
2	700	300	1	0	1	0	100%
3	1050	450	1	0	1	1	98%
2	700	300	1	1	0	0	98.5%
3	1050	450	1	1	0	1	96.3%
3	1050	450	1	1	1	0	96.6%
4	1400	600	1	1	1	1	95%

When comparing the results obtained from the proposed method with the results obtained from [7], whereby Alexnet archived an accuracy of 98.67%, trained on a larger dataset size. It is evident that increasing and using a larger PSB training set might yield better results greater than 95%.

C. *Lable and Text Classification*

Once the fine-tuning process is successful and the desired accuracy is archived, we verify that the model is fully functional by providing an input image of a scene. The image of choice as seen in Fig. 5, 6 and 7, were classified under their respective class label using fine-tuned AlexNet CCN. With reflection to our proposed method, we do not only want to classify which the class does the input image belong to, we also wanted to extract and classy the text presented. We were able to archive that, as illustrated on Fig. 5(a); 6(a) and 7(a) respectively using MSER algorithm to extract text regions and generate bounding boxes around wanted regions. We then use ORC to classify the selected regions and provide textual information represented on the images.

(a)MSER text region

(c) OCRtext verification

```
ans =
    'NO NOT DRINK
    PLEASE DON'T DUMP
    IN TOILETS
    HER\
    3ITY(
    'EY'BA
    2OUN(
    PUBLIC
    TOILETSA
```

(b) OCR extracted text

Fig. 5. Do Not Drink text extraction and verification

(a)MSER text regions

```
ans =
    'NO
     SMOKINGI
     [IL
     '
```

(b) OCR extracted text

(c) OCR text verification

Fig. 6. No Smoking text extraction and verification

(a)MSER text regions

(c) OCR text verification

```
ans =
    'Animals Allowed.
     ...it's the law!
    (CA. Penal Code 365.73)
    EXCEPTIONS:
    DOG FOR THE BLIND = SIGNAL DOG FOR THE DE.
    RVICE DOG FORYHE DISABLED/HANDICAPPED
    RVICE DOG FRAUD IS A CRIME
    rvy nemuw |)OG5"av: nan ADA Servlu Dogs

    Lelill
    Sorry.
    .
```

(b)OCR extracted text

Fig. 7. No Pets text extraction and verification

The sample verification images representing a scene have produced results that are in line with our primary objective, which was to classify images of a scene based on classes contained in our PSB dataset and produce textual

information contained in the image. The textual information was obtained successfully but with error as some words were misspelled and some missing characters, as illustrated on Fig. 5(b), 6(b) and 7(b) respectively. It was also necessary to verify that the key words which provides meaning within the image was present, as illustrated on Fig. 5(c), 6(c) and 7(c) respectively. From the sample data used, the results show that some textual data can provide same meaning of information relating to the symbol represented on the verification image but fail to contain the key words represented by class labels.

V. CONLUSION

This paper summarizes the benefiting of using convolutional neural networks for the purpose of image classification, and the computer vision text extraction and classification methods. The purpose was to classify four different classes of PSB, and extract the textual information, in order to aid visually impaired person by providing warnings of unsafe environments. The proposed method achieved remarkable results, which includes a 95% accuracy on our 2000 PSB dataset. The text extraction and classification results were acceptable as most of the sample verification images were able to extract and classify the textual information successfully. Future work includes comparing GoogLeNet's and ResNet's performance to AlexNet on our current PSB dataset. In [15] it has been stated that GoogLeNet can outperform any traditional CCN model, but it is also mentioned that it performs best on large size datasets; it would be of interest to investigate. We aim to increasing the PSB dataset and evaluate the performance in comparison with other state-of-the-art CNNs. The errors on text extraction and classification will also be addressed.

ACKNOWLEDGMENT

This work is supported by the National Research Foundation of South Africa (Grant Number 99383).

REFERENCES

[1] K. Motshoane, C. Tu, P.A. Owolawi, "An Overview of Vision Substitution System for the Visually Impaired," in Proc. of the 2017 3rd IEEE Int. Conf. on Computer and Communications, ICCC 2017, 13-16 December 2017, Chendu, China.

[2] E. Salahat, M. Qasaimeh, "Recent Advances in Features Extraction and Description Algorithms: A Comprehensive Survey." in Proc. of the 2017 IEEE Int. Conf. on Industrial Technology, ICIT 2017, 22-25 March 2017, Toronto, ON, Canada.

[3] N. Kajbakhsh, J. Y. Shin, S. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang. "Convolutional neural networks formedical image analysis: Full training or fine tuning?," IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp 1299-1312, 2016.

[4] A. Krizhevsky, I. Sutskever, G.E. Hinton, "ImageNet Classification with Deep Convolutional Neural Network", Communication of the ACM, vol. 60, no. 6, pp. 84–90, 2017.

[5] C. Szegedy, D. Anguelov, D. Erhan, V. Vanhoucke, W. Liu, Y. Jia, P. Sermanet, S. Reed, A. Rabinovich,"Going Deeper with Convolutions," in Proc. of the 2015 Computer Vision and Pattern Recognision, CVPR 2015, 8-10 June 2015, Boston, Massachusetts.

[6] K. He, X. Zhang, S. Ren, J. Sun "Deep Residual Learning for Image Recognition" in Proc. of the 2015 Computer Vision and Pattern Recognision, CVPR 2015, 8-10 June 2015, Boston, Massachusetts.

[7] M.S. Elmahdy, S.S Abdeldayem, and I.A. Yassine," Low Quality Dermal Image Classification Using Transfer Learning," in Proc. of the 2017 IEEE EMBS Int. Conf. on Biomedical & Health Informatics, ICBHI 2017, 16-19 February 2017, Orlando, FL, USA.

[8] A. Cordova-Cruzatty, B.M.D. Barreno, B.J.M J´acome, "Precise Weed and Maize Classification through Convolutional Neuronal Networks." in Proc. of the 2017 IEEE Second Ecuador Technical Chapters Meeting, ETCM 2017, 16-20 October 2017, Salinas, Ecuador.

[9] S. Gu, L. Ding, Y. Yang, X. Chen," A New Deep Learning Method Based on AlexNet Model and SSD Model for Tennis Ball Recognition," in Proc. of the 2017 IEEE 10th Int. Workshop on Computational Intelligence and Application, IWCIA 2017, 11-12 November 2017, Hiroshima, Japan.

[10] M.N. Stolar, M. Lech, R.S. Bolia, M. Skinner," Real Time Speech Emotion Recognition Using RGB Image Classification and Transfer Learning," in Proc. of the 2017 11th Int. Conf. on Signal Processing and Communication Systems, ICSPCS 2017, 13-15 December 2017, Gold Coast, QLD, Australia.

[11] R. Swathika and T.S. Sharmila,"Emergency exit sign detection system for visually impaired people," in Proc. of the 2016 Int. Conf. on Inventive Computation Technology, ICICT 2016, 26-27 August 2016, Coimbatore, India Australia.

[12] R. Islam, C. Mondal, K. Azam, A. Syed, J. Islam,"Text Detection and Recognition Using Enhanced MSER Detection and a Novel OCR Technique," in Proc. of the 2016 5th Interanational Conference on Informatics, Electronics and Vision, ICIEV 2016, 13-14 May 2016, Dhaka, Bangladesh.

[13] X. Chen., L. Hong., G. Huang. "Ensemble Learning for Image Recognition," in Proc. of the 2017 12th Int. Conf. on Intelligent Systems and Knowledge Engineering, ISKE 2017, 24-26 November 2017, Nanjing, China.

[14] J.P.N. Acilo, A.G.S.D. Cruz, M.K.L. Kaw, M.D. Mabanta, V.G.G. Pineda, E. A. Roxas, "Ensemble Learning for Image Recognition," in Proc. of the 2018 IEEE 14th International Colloquium on Signal Processing & its Applications, CSPA 2018, 9 -10 March 2018, Penang, Malaysia.

[15] Z. Zhu, J. Li, L. Zhuo, J. Zhang. "Extreme Weather Recognition using a Novel Fine-Tuning Strategy and Optomized GoogLeNet," 2017 International Conference on Digital Image Computing Techniques and Application, DICTA 2017, 29 Novemver – 01 December 2017, Sydney,NWS, Australia.

Switch: A Battery Conservation Framework for Mobile Devices

Herman Barnardt and Marijke Coetzee
Academy of Computer Science and Software Engineering,
University of Johannesburg
Johannesburg, South Africa
hmbarnardt@gmail.com, marijkec@uj.ac.za

***Abstract*—Sophisticated mobile applications may require more resources than are readily available on mobile devices. Mobile device resources such as processing power and storage can be extended with cloud-based mobile augmentation. However, some resources, specifically battery life and bandwidth, cannot be augmented. This research identifies that it is important to be able to estimate the energy consumption of both offloaded and local tasks when making offloading decisions. Due to the fact that the energy consumption profile of mobile devices with diverse capabilities are not the same, this aspect needs to be considered. This research proposes the Switch framework to conserve the limited battery life on mobile devices using a device specific energy consumption profile. The evaluation of the framework suggests that Switch can successfully be used to conserve battery life on mobile devices by making intelligent offloading decisions.**

Keywords—offloading, mobile devices, decision making, battery life, cloud-based mobile augmentation

I. INTRODUCTION

Statistics show that around 7.7 billion mobile devices are currently being used throughout the world. The number is predicted to rise to 12.1 billion in 2018, where 2.1 billion of these devices are smart devices [1]. As many mobile applications have resource intensive operations it is not surprising that the biggest consumer problem with mobile devices is the lifespan of the battery [2], [3]. Mobile devices generally all suffer from limited CPU, memory, storage capacity, and battery life; where battery life has been identified as the most limiting factor [3]. Many other factors such as backlighting, wireless connections, and processor speed all have an impact on the energy consumption of the device. These limitations may cause the mobile operating system to ask the application to shut down or slow program execution. The focus of current research is to provide interventions that can optimize resource usage to ensure that an application can perform its task without any interruption [4 - 8]. Due to its inherent nature, battery life is the one resource of a mobile device that cannot be augmented. In this regard, a large body of research is being conducted into techniques and frameworks to conserve battery life by offloading computation to the cloud [6 - 11]. It is important to consider that available bandwidth has a large impact on the power consumption of a mobile device when performing offloading. Depending on bandwidth, energy consumption can be doubled in the presence of packet loss and increased propagation delay.

Current research in mobile cloud computing [12 – 15] focuses on techniques that can be used to support resource-demanding mobile applications. Offloading is defined as the process of moving a task from a resource poor client to a resource rich server such as the cloud [16]. Offloading from a mobile device to the cloud, is a challenging task to accomplish which requires an understanding of the requirements of the cloud server, the identification of offloadable components and data to be offloaded, are some of the factors that should be considered when creating a mobile cloud computing offloading framework. Research [17], [18] has shown that offloading can be used to conserve the battery life of mobile devices in the right circumstances. Before a process is executed, a decision has to be made to either offload the process, or execute the process on the mobile device.

The goal of this research is to create a framework that conserves battery life on mobile devices by making intelligent offloading decisions. To achieve this, offloading frameworks are discussed in Section II to identify the requirements that the proposed framework needs to address, shown in Section III. Section IV describes the Switch framework architecture and components and its decision making process. Section V gives the Switch prototype implementation and evaluation and finally, the paper is concluded.

II. RELATED WORK

Mobile devices can be augmented with resources made available by the cloud by offloading processes. To gain an understanding how offloading is performed, the most representative offloading frameworks that make use of dynamic decision-making namely CloneCloud [6] and MAUI [7] are discussed. From this discussion it becomes clear that these frameworks are complex to implement and require hardware components to measure the consumption of battery life as both employ an external power monitor, making them impractical for commercial use.

A. Offloading frameworks

1. MAUI

MAUI (Mobile Assistance Using Infrastructure) [7], is an offloading approach that focuses the optimization of energy consumption and execution times of mobile applications. MAUI enables the fine-grained energy-aware offload of

either mobile processes or mobile code elements without much programmer effort, as code annotations indicate which methods can be executed remotely. Thus applications need to be modified so that they can be offloaded. This approach is limited to the Microsoft .NET platform [19] which limits the type of applications that can be offloaded. First, application code is replicated and placed on the cloud server. Thereafter the MAUI system evaluates the code and serializes and profiles all methods to determine the offloading cost. The offloading cost is determined before the execution of a task that can be offloaded. Part of the MAUI solver is the profiler that determines the cost and decides whether to run the method locally or remotely. The profiler used by the MAUI solver takes three factors in to consideration when determining cost of offloading, namely (1) the device's energy consumption characteristics, (2) the program characteristics, such as the running time and resource needs of individual methods and (3) the network characteristics of the wireless environment, such as the bandwidth, latency, and packet loss. The offloading occurs over Wi-Fi or mobile networks.

2. CloneCloud

CloneCloud [6] removes the need to duplicate and rework a mobile application's code to execute on the cloud. CloneCloud creates a virtual machine of the device and runs the virtual machine on the cloud. Whenever a process is executed, the application moves the state of the device to the virtual machine and the execution continues from where the device stopped executing. A static analyser evaluates the methods of the application offline and the appropriate methods or parts of the code are marked so that they can be offloaded. If one of the marked methods is executed, a profiler decides whether or not the process should be offloaded. The profiler gathers the data regarding the execution of the method, and determines whether or not the method should be executed locally or be offloaded to the cloud. The profiler used in CloneCloud uses execution time and energy consumed by the mobile device to inform the decision. The application state is offloaded over both mobile networks and Wi-Fi. Code partitioning is static as the developer marks methods that can be offloaded during development. CloneCloud achieves cloud based mobile augmentation by creating virtual instances of a mobile device on the cloud. For the virtual instance of the device to execute the code from the physical device, the state of the physical device has to be transferred to the cloud. The amount data that specifies the state of the device, or the application, may be large and lengthen the time the method takes to execute or how much battery life is used while executing the method [6].

B. Challenges of offloading

To gain the advantages of offloading, certain challenges need to be overcome as follows:

- *Bandwidth available to the mobile device*: Cellular networks are highly pervasive but also have an associated cost and do not always have reliable bandwidths. In contrast, Wi-Fi can provide network access at greater bandwidth but does not cover such a large area as the mobile networks.
- *Network availability:* Network availability is critical as mobile devices may not be capable to complete complex tasks without offloading.
- *Network heterogeneity:* Different networks and communication protocols need to be considered when the offloading server is created.

III. FRAMEWORK REQUIREMENTS

To define an offloading framework that can conserve battery life when executing a task, the following requirements need to be addressed.

- *Make intelligent offloading decisions*: When offloading decisions are made, the decision making component must consider all factors that can influence the energy consumption of the mobile device, and choose the option that consumes less energy by using a device specific energy consumption profile that is scalable.
- *Multiple network support:* Both Wi-Fi and cellular networks should be supported as a Wi-Fi network may not always be available. Wi-Fi is the preferred network to use as it is more stable, has relatively high bandwidth, and consumes less power.
- *Be lightweight*: The solution should consume as little energy as possible and execute as quickly as possible, as lightweight processes will not have a detrimental impact on user experience or battery life.
- *Portable*: The solution should be deployable on any device and integrated into any app with minimal effort and without additional hardware.

IV. SWITCH FRAMEWORK

This research proposes a novel software profiler to inform offloading decisions of the energy consumption of a mobile device. For this purpose, the Switch framework is presented as a solution to the problem of limited battery life on mobile devices. Switch can assist developers to determine whether or not a task should be offloaded to the cloud in order to reduce energy consumption so that the battery life of a mobile device is conserved. The goal is achieved by developing both a software profiler, which estimates the amount of battery life needed when executing a task, and a decision making component that uses the estimates provided by the profiler. Before a process is executed either locally or in the cloud, a comparison is done between the expected energy consumption when executing the process locally and the expected energy consumption when offloading the process. The process is executed where the least amount of battery life is consumed.

Offloading frameworks either employ virtual machine cloning or code offloading. This research uses code offloading where a function is executed on a cloud server instead of the local method, similar to what MAUI does. The offloading decision is influenced by the size of data, bandwidth, communication protocol, code complexity and time to execute locally, which are difficult to determine. Here, a device specific energy consumption profile characterizes these factors. As a hardware based profiler is

labour-intensive and not scalable, this research makes use of a software based approach that does not require expensive external power meters and enables the monitoring of the application on various levels of granularity.

A. Switch architecture

The contribution of Switch is to support developers to define offloading decisions for applications without requiring the integration of a large framework in both the app and the cloud. In order to achieve this, the energy consumption of the mobile device needs to be defined, offloadable tasks identified, and the duration of local execution and the size of the data communicated when offloading should be measured.

The architecture of Switch, shown in Figure 1, consists of a number of components. A mobile app is installed on the mobile operating system. From within the mobile app, the Switch component is called, consisting of a Decision Making component, a Profiler component and an Energy Consumption Profile component. First, before offloading decisions can be made, the energy consumption profile model is determined for the specific mobile device. A number of tests are run for each network type the device is connected to. Similar sets of tests are executed for local execution. Historical averages of the local and cloud execution costs are stored. The results are examined and the model is created. The creation of the energy consumption profile is shown by the communication between the Profiler and the Energy Consumption Profile in figure 1 (a).

Next, a mobile app developer defines a mobile app with any number of methods. Offloadable methods are identified and their server-side counterparts are created in the cloud. Traditionally, the app developer would decide whether or not the method is to be offloaded. However, by integrating Switch, such decisions are made on a per execution basis. The communication between the mobile app and Switch is shown in figure 1 (b).

Within Switch, the Decision Making component makes the offloading decision by utilising the Profiler component and the Energy Consumption Profile, shown in figure 1 (c) and (d). The Profiler component continuously monitors the state of the mobile device and the app. The Energy Consumption Profile component contain models that are used to estimate the energy consumption when provided with a set of variables. The estimated costs determine the outcome of the decision made by the decision making component. If the decision is made to offload the app communicates with the cloud, shown in figure 1 (e).

Switch is used to make the offloading decision for mobile apps that use offloading. The Switch framework is integrated into existing applications without interfering with their execution or communication. The profiler gathers data regarding the device and makes the offloading decision.

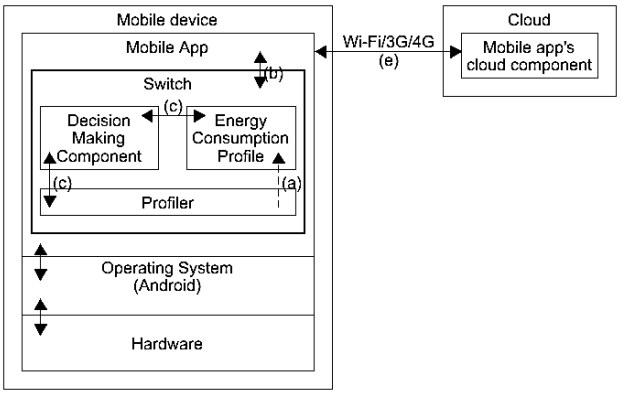


Figure 1: Switch framework architecture

B. Switch energy consumption profile

An energy consumption profile is defined that can support accurate estimations of battery life usage per mobile device. The factors that influences energy consumption should be continuously monitored to ensure the decisions made are correctly at any point in time. To be lightweight, the energy consumption profile should use a simple model to estimate energy consumption. An energy consumption profile is created from measurements taken during the execution of different tasks on a specific mobile device, such as executing a computationally intensive process for a set amount of time, or downloading a file over a specific network. An energy consumption profile per device per network enables an accurate estimation of battery life usage when communicating over a specific network.

The energy consumption profile used by Switch is described by four linear models that represent battery life usage. The models are used to estimate battery life consumption for a process or method by monitoring the relevant hardware components of the mobile device. Such measurements are more flexible as they are not bound to a specific mobile app. A monitoring application that observes the operating system is used to sample the energy consumption of the specific devices. The four models are defined as follows:

1. The energy consumption when the CPU of the mobile device is in use.
   $EC(CPU) = localRate(time) + localConstant$

localRate and *localConstant* are device specific constants. *time* is the length of time the CPU will be active.

2. The energy consumption when communicating with the 3G cellular network
   $EC(3G) = threeGRate(time) + threeGConstant$

threeGRate and *threeGConstant* are device specific constants. *time* is the length of time the 3G network radio will be active

3. The energy consumption when communicating with the 4G cellular network.
   $EC(4G) = fourGRate(time) + fourGConstant$

Where *fourGRate* and *fourGConstant* are device specific constants. *time* is the length of time the 4G network radio will be active

4. The energy consumption when communicating with Wi-Fi.

$$EC(WiFi) = wifiRate(time) + wifiConstant$$

Where *wifiRate* and *wifiConstant* are device specific constants. *time* is the length of time the Wi-Fi radio will be active.

Due to the energy consumption required to analyse the automatically gathered data, the task of generating the consumption model should be performed remotely on the cloud.

C. Switch decision making process

The offloading decision is calculated by comparing the estimated energy consumption when executing locally or when offloading to the cloud. Energy consumption is defined by two categories namely *communication*, which is determined by the size of the data, the bandwidth and the communication protocol used and *computation*, which is determined by code complexity and the time needed to execute.

$$shouldOffload = E_{local}(time) > E_{offloading}(bandwidth, size)$$

Where *shouldOffload* is a true or false value that determines whether or not offloading should occur. E_{local} represent the calculation of the cost of local execution. $E_{offloading}$ represents the calculation of the cost of offloading, described next.

1. *Cost of local execution*

$$localCost = localRate(time) + localConstant$$

Where *localRate* and *localConstant* are the constants provided by the CPU energy consumption model. *time* is the duration of local execution

2. *Cost of offloading*

$$offloadingCost = networkRate\left(\frac{size}{bandwidth}\right) + networkConstant$$

Where *networkRate* and *networkConstant* are constants provided by the energy consumption models. The network model used is determined by the network the device is currently connected to. *size* represents the size of the data that is transferred from and to the cloud. b*andwidth* is the current available bandwidth.

This research follows an approach to offload code to the cloud if energy can be saved, even if it may not result in a major saving. Another approach would be to enable the use of offloading if certain battery thresholds per device are passed.

V. Implementation and Evaluation

The Switch prototype is implemented on a Samsung Galaxy S7 Edge, SM-G900F, mobile device running Android 6.0.1 with a 3600mAh battery. The Network Signal Info Pro app, developed by KAIBITS Software [20] is used to get accurate readings of the networks the device is connected to. The GSam Battery Monitor app [21] developed by GSam Labs is used to monitor the battery life usage and the length of computation time. These third-party apps have been chosen out of a large number of contenders, due to their accuracy and usefulness, after a careful review by the researcher.

Two types of applications are used in the evaluation namely a *steganography application* which is a *c*omputationally intensive task that requires the transfer of large amounts of data when offloaded and a *prime number generation* application which is a computationally intensive task that requires very little data to be offloaded. The mobile apps that are used to evaluate the prototype offload to a Heroku container [22].The Heroku container used for the evaluation is on the free tier, it has 512MB of memory, and has 1x CPU share and between 1x and 4x compute share [23]. Although the resources available in the container used for offloading is not limitless it does surpass the available resources on the mobile device.

As mentioned, Switch is evaluated for both local computation and offloading over different types of networks where the estimated energy consumption is compared to the actual energy consumption. When estimating the energy consumption for offloading, the available bandwidth is measured and used to estimate the length of time of communication. The length of time of communication is used to estimate the energy consumption. The available bandwidth is used instead of the signal strength, because the signal strength is not a reliable indicator of how long communication lasts. To determine the estimation of energy consumption for the execution of a local task, the length of time of computation is measured.

1. Task 1 – Steganographic file encoding

The first task used is the steganographic encoding a 1.4 MB file into a 5 MB image. Offloading this task requires the image and the file to be uploaded and the encoded image to be downloaded. The task is executed locally, over Wi-Fi, 4G and 3G. The data collected from the executions of this task is shown in table 1. The estimated cost of task execution is compared to the actual cost that was measured.

TABLE 1 STEGANOGRAPHIC FILE ENCODING

	Time	**Estimated battery life**	**Measured battery life**
Local	7.291 seconds	0.057%	0.058%
3G	51 seconds	0.1768%	0.177%
4G	34 seconds	0.0983%	0.098%
Wi-Fi	104 seconds	0.2578%	0.258%

The large amount of data required to be transferred skews this task to be executed locally. It only takes 7.291 seconds to execute the task and consumes 0.058% of the battery life. As can be expected, it takes much longer to offload a task using a Wi-Fi connection and this consumes

more battery life. Even though a task executed using a 4G connection takes 34 seconds to complete, it is still much slower than a local execution and consumes more battery life. From the results it is clear that the estimated energy consumption is relatively accurate when compared to the actual battery life as measured by software.

2. Task 2 – Steganographic file decoding

The second task is to decode a 5 MB image. The encoded file contains a 1.4 MB file. Offloading this task requires the image to be uploaded and the encoded file to be downloaded. The task is executed locally, over Wi-Fi, 4G and 3G. The data collected from the executions of this task is shown in table 2. Comparing the results of this tasks with the results from the first task shows that in each instance this task consumes less battery life. The decrease in battery life consumption can be attributed to the decrease in the amount of data that is required to be transferred from 11.4 MB to 6.4 MB.

TABLE 2 STEGANOGRAPHIC FILE DECODING

	Time	Estimated battery life	Measured battery life
Local	2.304 seconds	0.0011%	0.001%
3G	30 seconds	0.1825%	0.183%
4G	10 seconds	0.075%	0.076%
Wi-Fi	114 seconds	0.1846%	0.185%

As with the first task the amount data that is required to be transferred skews the task to be executed locally. Local execution takes 2.304 seconds and consumes 0.001% of the battery life. The fastest alternative when offloading, 4G, takes 10 seconds and consumes 0.076% battery life. Due to the lower bandwidth available on Wi-Fi during testing the execution is even slower, 114 seconds, and consumes much more battery life, 0.185%.

3. Task 3 – Prime number counting

The third task used to evaluate Switch is calculating the number of primes there are between zero and ten million (10 000 000). Offloading this tasks requires the upload of an integer (4 B) and the download of an integer (4 B).

TABLE 3 PRIME NUMBER COUNTING RESULTS

	Time	Estimated battery life	Measured battery life
Local	31 seconds	0.0317%	0.032%
3G	16 seconds	0.0293%	0.03%
4G	12 seconds	0.0207%	0.02%
Wi-Fi	10 seconds	0.0127%	0.013%

The amount data to be transferred for this task is extremely small and the task is computationally complex, this skews the task toward offloading. Due to the complexity of the task execution on the mobile device takes 31 seconds and consumes 0.032% battery life. The battery life consumed when offloading over 4G is 0.02% and only takes 12 seconds. Over Wi-Fi the same operation takes 10 seconds and consumes 0.013% battery life.

A. Evaluation

When comparing the estimated cost for local execution with the cost when offloading for each of the networks, for the first and second tasks, the prototype suggests that local execution conserves battery life. Manual inspection of the measured energy consumption shows that it is the least expensive option. For the third task, the prototype suggests offloading in all cases when comparing the energy consumption estimates. Again, manual inspection of the measured energy consumption values, shows that local execution is the least expensive option.

For each of the models the average percentage error is calculated. The percentage error is used to show the difference between estimated values and measured values using [24]:

$$percent\ error = \frac{|measured - estimated|}{measured} \times 100$$

The inverse of the percentage error shows the accuracy of the model using:

$$accuracy\ percent = 100 - percent\ error$$

The calculation of the percentage errors for the CPU energy consumption model is shown in table 4.

TABLE 4 PERCENTAGE ERROR CALCULATION - CPU MODEL

	Estimated cost	Measured cost	Percentage error
Task 1	0.057%	0.058%	1.724137931%
Task 2	0.0011%	0.001%	10%
Task 3	0.0317%	0.032%	0.9375%

Averaging the percentage errors results in an average percentage error of 4.22%. The average percentage error shows that the CPU energy consumption model has an accuracy percentage of 95.78%.

Similarly, the percentage error for the 3G energy consumption model is 99.09, the 4G energy consumption model is 98.29% accurate and the Wi-Fi energy consumption model is 99.13%. The high accuracy percentages of the models show that the costs estimated by Switch are very close to the measured values.

VI. CONCLUSION

The energy consumption of a mobile device or an app can be measured using either a hardware or software-based approach where hardware based approaches result in profiles that are highly accurate but their creation are labour-intensive and not scalable. Software based approaches result in less accurate profiles but do enable the monitoring of the application on many different granularities and do not require expensive external power meters.

Previous research required specialized hardware to function or to measure energy consumption, or required in depth analysis of the methods executed by the app that the prototypes were integrated with. The Switch framework takes a different approach in that it aims to be integrated with any mobile app that uses offloading and software to measure energy consumption.

The framework focuses on a device specific energy consumption profile and not on the energy consumption of

the mobile app. When the energy consumption profile is created for a device, minimal measurements are required to be made by the developer. As verified in the prototype, the battery life of a mobile device can be conserved by integrating the prototype into an app.

REFERENCES

[1] S. Radicati, 'Mobile Statistics Report, 2014-2018 (Executive Summary)', The Radicati Group, Inc., Palo Alto, CA, USA, Feb. 2014.

[2] D. Ferreira, A. K. Dey, and V. Kostakos, 'Understanding Human-Smartphone Concerns: A Study of Battery Life', in *Pervasive Computing*, San Francisco, USA, 2011, pp. 19–33.

[3] B. Reed, 'Apple might want to start taking gripes about the iPhone's battery life seriously', *BGR*, 14-Aug-2014.

[4] S. Abolfazli, Z. Sanaei, E. Ahmed, A. Gani, and R. Buyya, 'Cloud-Based Augmentation for Mobile Devices: Motivation, Taxonomies, and Open Challenges', *IEEE Commun. Surv. Tutor.*, vol. 16, no. 1, pp. 337–368, 2014.

[5] S. Abolfazli, Z. Sanaei, M. Shiraz, and A. Gani, 'MOMCC: Market-oriented architecture for Mobile Cloud Computing based on Service Oriented Architecture', in *2012 1st IEEE International Conference on Communications in China Workshops (ICCC)*, 2012, pp. 8–13.

[6] B.-G. Chun, S. Ihm, P. Maniatis, M. Naik, and A. Patti, 'CloneCloud: Elastic Execution Between Mobile Device and Cloud', in *Proceedings of the Sixth Conference on Computer Systems*, New York, NY, USA, 2011, pp. 301–314.

[7] E. Cuervo *et al.*, 'MAUI: Making Smartphones Last Longer with Code Offload', in *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services*, New York, NY, USA, 2010, pp. 49–62.

[8] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, 'The Case for VM-based Cloudlets in Mobile Computing', *IEEE Pervasive Comput.*, vol. Early Access Online, 2011.

[9] R. Kemp, N. Palmer, T. Kielmann, and H. Bal, 'Cuckoo: A Computation Offloading Framework for Smartphones', in *Mobile Computing, Applications, and Services*, M. Gris and G. Yang, Eds. Springer Berlin Heidelberg, 2012, pp. 59–79.

[10] J. Parkkila and J. Porras, 'Improving battery life and performance of mobile devices with cyber foraging', in *2011 IEEE 22nd International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, 2011, pp. 91–95.

[11] A. Yousafzai, A. Gani, R. Md Noor, A. Naveed, R. W. Ahmad, and V. Chang, 'Computational offloading mechanism for native and android runtime based mobile applications', *J. Syst. Softw.*, vol. 121, pp. 28–39, Nov. 2016.

[12] H. T. Dinh, C. Lee, D. Niyato, and P. Wang, 'A survey of mobile cloud computing: architecture, applications, and approaches', *Wirel. Commun. Mob. Comput.*, vol. 13, no. 18, pp. 1587–1611, 2013.

[13] M. Alizadeh and W. H. Hassan, 'Challenges and opportunities of Mobile Cloud Computing', in *Wireless Communications and Mobile Computing Conference (IWCMC), 2013 9th International*, 2013, pp. 660–666.

[14] N. Fernando, S. W. Loke, and W. Rahayu, 'Mobile cloud computing: A survey', *Future Gener. Comput. Syst.*, vol. 29, no. 1, pp. 84–106, Jan. 2013.

[15] M. Souppaya and S. Scarfone, 'Guidelines for Managing the Security of Mobile Devices in the Enterprise', National Institute of Standards and Technology (NIST), SP 800-124 r1, Jun. 2013.

[16] K. Kumar, J. Liu, Y.-H. Lu, and B. Bhargava, 'A Survey of Computation Offloading for Mobile Systems', *Mob. Netw. Appl.*, vol. 18, no. 1, pp. 129–140, Feb. 2013.

[17] M. V. Barbera, S. Kosta, A. Mei, and J. STEFA, 'To offload or not to offload? The bandwidth and energy costs of mobile cloud computing', in *2013 Proceedings IEEE INFOCOM*, 2013, pp. 1285–1293.

[18] K. Kumar and Y.-H. Lu, 'Cloud Computing for Mobile Users: Can Offloading Computation Save Energy?', *Computer*, vol. 43, no. 4, pp. 51–56, Apr. 2010.

[19] Microsoft, '.NET - Powerful Open Source Cross Platform Development', *Microsoft*, 2016. [Online]. Available: https://www.microsoft.com/net. [Accessed: 22-Nov-2016].

[20] KAIBITS, 2015. KAIBITS Software. KAIBITS. Available: http://www.kaibits-software.com/ (Accessed 10 August 2017).

[21] GSam Labs, 2013. GSam Labs. GSam Labs. Available: http://www.gsamlabs.com (Accessed 10 August 2015).

[22] Heroku, 'About Heroku | Heroku', 2017. [Online]. Available: https://www.heroku.com/about. [Accessed: 08-Dec-2017].

[23] Heroku, 'Dyno Types | Heroku Dev Center', 2017. [Online]. Available: https://devcenter.heroku.com/articles/dyno-types. [Accessed: 08-Dec-2017].

[24] T. Helmenstine, 'Calculate Percent Error', *Science Notes and Projects*, 14-Sep-2017. .

Gamification in m-Health Applications and its Effects on Health Behavioural Changes in Net Generation Students

Jason Goldhill
Department of Information Systems
University of Cape Town
Cape Town, South Africa
gldjas005@myuct.ac.za

Sumarie Roodt
Department of Information Systems
University of Cape Town
Cape Town, South Africa
sumarie.roodt@uct.ac.za

***Abstract*—This research examines Gamification in m-Health Applications and its Effects on Health Behavioural Changes in Students. The four central themes highlighted in this research are Gamification, m-Health Applications, Persuasive Computing, and the Net Generation. Gamification is a relatively new topic that is beginning to be looked seriously as a way to overcome problems and improve processes due to its characteristics. One of these characteristics is its ability to persuade and motivate its users, which can have positive effects on their behaviours and attitudes. The Net Generation has been shown to be particularly susceptible to Gamification. Therefore, it is hoped that Gamification, when combined with the ever-increasing access to and improvement of mobile technology, could provide a serious solution to some of the health problems and behaviours common to the Net Generation. This research surveyed members of the Net Generation from the University of Cape Town in order to determine how their use of Gamification within m-health applications had affected their health behaviours. The findings, while not statistically representative of the entire Net Generation, show that there is evidence that the inclusion of Gamification in m-health applications has a positive effect on the health behaviours of members of the Net Generation. Although more research needs to be conducted to confirm this, the findings of this research imply that Gamification, when combined with m-health application, could indeed provide a solution to some of the health problems and behaviours common to the Net Generation.**

Keywords—Digital Natives, Gamification, Generation Y, Health Behaviours, m-Health Applications, Millennials, Net Generation, Persuasive Computing, Students

I. Introduction

Diseases and negative health habits are currently major issues for many people in the world, particularly young adults and children who are part of the Net Generation. The only way for many of these patients to improve their health is by changing their health behaviours and making "long term lifestyle changes" [1]. However, according to [1], the Net Generation is developing shorter attention spans than previous generations, which is causing them to form negative health habits and is making it more difficult for them to change their health behaviours.

Access and affordability of healthcare are another large barrier to changing one's health behaviours, however mobile phone technology and usage are climbing greatly while their costs are falling [2]. Mobile phones could therefore potentially begin to break down the barriers of access and affordability of healthcare.

Gamification is becoming an increasingly popular subject and has been proven to produce results [3]. Multiple tests and empirical studies have been conducted with gamification in different fields and sectors, particularly education, which have shown success [3][4]. However there seems to be a lack of research conducted on gamification in the health sector [5][6]. [5][6] believe that there is a lot of promise for the use of gamification in the health sector, particularly in conjunction with m-health applications.

There are no efficient and affordable methods of improving health behaviours of members of the Net Generation that are engaging enough to hold their attention. It is unclear whether the combination of gamification and m-health applications will be a viable solution to this problem.

The purpose of this research is to begin to fill in the gap regarding the use of gamification in the health sector, particularly m-health applications, in the hope that it will begin to shed some light on possible solutions to changing individuals' health behaviours.

This research will contribute further information on the relationships between gamification and health behaviour changes that will be useful for both designers and users of m-health applications. By taking this information into consideration when making decisions about the design or use of m-health applications both designers and users will be able to make more informed decisions that will hopefully lead to more positive effects on health behaviours.

II. Literature Review

A. Gamification

"Gamification is the use of game design elements in non-game contexts" [7][8][9][10][11]. This definition, first produced by [7], has become widely used and accepted as the standard definition of gamification, as evidenced by the number of papers that use it. However, one other definition of gamification does exist [7][9]. [12] attempt to define gamification from the perspective of service marketing. "Gamification refers to: a process of enhancing a service with affordances for gameful experiences in order to support user's overall value creation" (Huotari & Hamari, 2012, p. 19). According to Huotari and Hamari (2012) their definition differs from the definition of Deterding et al. (2011) as it emphasizes the goal of gamification – improving a service – rather than the methods of gamification.

Both Deterding et al. (2011) and Groh (2012) state that in order for the definition of Deterding et al. (2011) to have meaning, one must first understand the definitions of the parts that make it up, namely: Game, Element, Design, and Non-Game Contexts.

B. M-Health Applications

e-Health (or electronic health) is healthcare that, in some form, is supported by electronic means (Liu, Zhu, Holroyd, &

Seng, 2011)[13]. m-Health (or mobile health) is a subset of e-health that makes specific use of mobile technologies, such as smartphones, tablets, the Internet, etc. (Free et al., 2013; Liu et al., 2011)[13][14]. In other words, m-health refers to "mobile computing, medical sensor, and communications technologies for health care" (Liu et al., 2011, p. 2022; Wu, Carpenter, & Himes, 2015, p. 1)[13][15]. m-Health applications are pieces of software that run on a mobile device, either natively or on the web, through which a user accesses an m-health technology. m-Health applications offer a wide "range of functions from clinical decision support systems and data collection tools for healthcare professionals, to supporting health behaviour change and chronic disease management by patients in the community" (Free et al., 2010)[16].

C. Net Generation

The Net Generation (also referred to as Generation Y, Millennials, or Digital Natives) is the generation of people who have grown up surrounded by technology such as computers, TV, smartphones, video games, and the Internet (Jones, Ramanau, Cross, & Healing, 2010; Leung, 2004; Oblinger, Oblinger, & Lippincott, 2005; Sandars & Morrison, 2007)[17][18][19][20]. Although there is general consensus as to what the Net Generation is, there seems to be some disagreement as to exactly who it includes. Leung (2004) [18] states that it is anyone born between 1977 and 1997, while Sandars and Morrison (2007)[20] shorten that period considerably to between 1982 and 1991. Jones et al. (2010)[17] simply state that it is anyone born since digital technology began to become commonplace in the 1980s. Due to the fact that the Net Generation has grown up surrounded by technology they have developed a natural talent for using it (Jones et al., 2010)[17]. The Net Generation are severely different from previous generations as they have become reliant on and comfortable with the technology that they have grown up with (Jones et al., 2010; Sandars & Morrison, 2007)[17][20]. Jones et al. (2010)[17] provide an amusing, yet accurate, quote that says the Net Generation is "a generation who think IM, text and Google are verbs not applications!" (p. 722), which sums up this new reliance on and comfort with technology. Their reliance on technology has vastly altered the way the Net Generation "learn, shop, work, play and communicate" (Sandars & Morrison, 2007, p. 85)[20].

D. Persuasive Computing

Persuasive computing refers to the ability of computers and technology to change the attitudes and behaviours of users who interact with it (Berdichevsky & Neuenschwander, 1999; Davis, 2009; Fogg, 1998, 1999)[21][22][23][24]. Persuasive computing is the intersection of two fields of study, persuasion and computers, and the study of this intersection is called captology (Davis, 2009; Fogg, 1998, 1999)[22][23][24]. According to Fogg (1999)[24], it is important to determine the intentionality of the persuasion of computers and technologies. Persuasive computing only refers to computers and technologies that are "intentionally designed [or used] to change a person's attitudes or behaviour in a predetermined way" (Fogg, 1999, p. 27)[24]. Lockton, Harrison and Stanton refer to this as "Design with Intent" and state that "Design with Intent" is a requirement for a technology to be considered persuasive. A change in attitudes or behaviour resulting from the use of a computer or technology that was not intended is considered a side effect, and not persuasive computing (Fogg, 1998, 1999; Lockton, Harrison, & Stanton, 2008) [23][24][25]. Fogg (1998) [23] proposes three types of intent that could be used to define a technology as persuasive – endogenous, exogenous, and autogenous.

III. Research Design

The selected philosophy for this research was positivism as it was believed that there is an "observable social reality" (Saunders et al., 2009, p. 113) [26] regarding the relationship between the use of gamified m-health applications and health behaviour changes in students, and that this relationship can be measured and facts can be drawn from this.

For the purpose of this research, the deductive approach was taken as the research was conducted under the hypothesis that there is a relationship between using gamified m-health applications and behavioural changes in students, and the purpose of this research is to either prove or disprove that hypothesis.

The purpose of this research is explanatory as it is trying to determine the relationship between the use of gamified m-health applications and behavioural changes in students.

Although an experiment would have been an acceptable strategy for this research, due to time constraints and the complexities involved in running an experiment, a survey strategy was used instead. The survey strategy still provided a way to study the relationships between variables (in the case of this research, the use of gamified m-health applications and behavioural changes in students) in a far more time efficient and less complex manner than an experiment.

A cross-sectional timeframe was used for this research as this research aimed to determine the current relationship between the use of gamified m-health applications and behavioural changes in students. As previously mentioned, this research was also subject to time constraints, and thus a cross-sectional timeframe was more feasible.

In order to focus on the key themes, a sample of students from the University of Cape Town was used as they are likely to have had access to gamified m-health applications and are students, particularly members of the Net Generation. As the selection of non-probability samples are also based on the subjective judgement of the researcher (Saunders et al., 2009)[26], the INF1002S (Information Systems I) class was initially selected as the sample for this research. There were several reasons for this selection. INF1002S is a compulsory class for all first year Bachelor of Commerce and Bachelor of Business Science students, no matter their planned field of specialization, which means it provided one of the largest and most diverse samples in a single class at the university. Access to an Information Systems class also proved to be easier to obtain compared to other classes as this research was being conducted for the Department of Information Systems.

Due to a low response rate from the selected sample of the INF1002S class (see section 4.7.2. for more information on this), the sample was expanded to include members of the INF4027W class as well in order to increase the total number of responses. The INF4027W class still forms part of the target population as they are members of the Net Generation. The reasoning behind the selection of the INF4027W class was simply ease of access, with the researcher being a member of the class and thus having direct access to the students.

For the purposes of this research, only closed questions were asked as the nature of the research required that respondents' answers did not vary excessively if quantitative analysis was to be performed. Out of the possible closed questions, only list, category, and rating questions were asked as the data that this research was looking to gather was predominantly behaviour (category questions) and opinion (rating questions) based. Questions included an 'Other (Please specify)' option where it was difficult to provide an exhaustive list of options. In order to improve the response rate, the questionnaire was accompanied by a cover letter detailing the background of the research and was designed in a clear and appealing manner (Saunders et al., 2009)[26].

IV. Data Analysis

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

A. General Information

The first question of the research instrument asked users whether they currently use or have ever used m-health applications on their mobile phones or tablets. This question was used to categorise respondents as either Users or Non-Users. The results of this can be seen in Table 1.

Table 1: Total Users vs. Non-Users

	Users	Non-Users	TOTAL
Count	20	16	36

Table 1 shows that 20 out of the 36 respondents (56%) either currently use or have used m-health applications in some way. This shows that more than half of the sample are aware of m-health applications and have actively pursued their use.

The 20 respondents who were categorised as Users were asked whether the m-health applications they had used contained any gamification elements. The results of this can be seen in Table 2.

Table 2: Level of Gamification Incorporation in Used m-Health Applications

	Contained Gamification Elements	Did Not Contain Gamification Elements	TOTAL
Count	13	7	20

Table 2 shows that 13 out of the 20 respondents (65%) who use or have used m-health applications stated that the applications contained gamification elements. This could mean one of two things. Either members of the sample (the Net Generation) are more inclined to use m-health applications that contain gamification elements (explored in the third secondary research question), or the majority of m-health applications contain gamification elements.

The results displayed in Tables 1 and 2 show both that there are a significant number of users of m-health applications and that a significant number of these applications contain gamification elements. Therefore, there is enough data available to answer the research questions.

B. Main Research Question

The main research question asked how the inclusion of gamification elements in m-health applications affects health behaviour changes in students. In order to answer this question it must be determined whether there is correlation between the inclusion of gamification elements in used m-health applications and perceived health behaviour changes in students. The collected data outlining the relationship between the level of agreement that health behaviours changed and whether the used m-health applications contained gamification elements can be seen in Table 3, while the expected values should there be no correlation can be seen in Table 4.

Table 3: Collected Responses of Level of Agreement that Health Behaviours Changed vs Whether the Used Applications Contained Gamification Elements

Level of Agreement that Health Behaviours Changed	Contained Gamification Elements	Did Not Contain Gamification Elements	TOTAL
Strongly Disagree	1	1	2
Disagree	2	2	4
Neutral	4	2	6
Agree	4	0	4
Strongly Agree	2	2	4
TOTAL	13	7	20

Table 4: Expected Responses of Level of Agreement that Health Behaviours Changed vs Whether the Used Applications Contained Gamification Elements Under the Assumption That There is No Correlation

Level of Agreement that Health Behaviours Changed	Contained Gamification Elements	Did Not Contain Gamification Elements	TOTAL
Strongly Disagree	1.3	0.7	2
Disagree	2.6	1.4	4
Neutral	3.9	2.1	6
Agree	2.6	1.4	4
Strongly Agree	2.6	1.4	4
TOTAL	13	7	20

Based on the collected responses in Table 3 and the expected responses in Table 4, the chi-squared test results in a p-value of 0.53. In other words, there is a 53% chance that the collected data or data more extreme could occur simply by chance. The collected data is therefore not statistically significant and no statistical conclusions about the entire population can be made based on it.

The counts provided by Table 3 can then be transferred to a graphic format, shown in Figure 1.

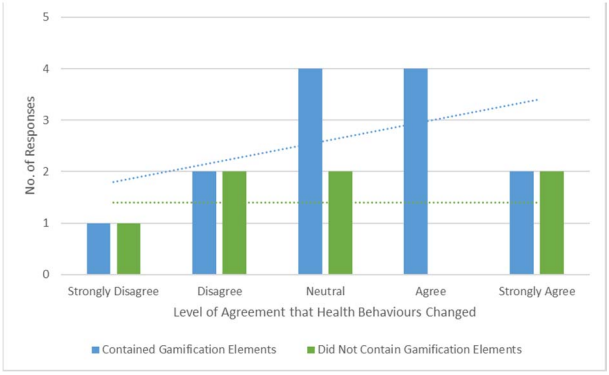


Figure 1: Levels of Agreement that Health Behaviours Changed by Whether the Used Applications Contained Gamification Elements or Not

In Figure 1, the number of responses per level of agreement that health behaviours changed is represented by the blue series for responses that claimed the used app contained gamification elements and the green series for responses that claimed the used app did not contain gamification elements. Each series has then had a trend line drawn in. These trend lines represent the observed relationship between whether the used app contained gamification elements and how strongly the respondents perceived their health behaviours changed. In Figure 1, the blue trend line has a positive gradient and thus shows that there is a positive relationship between the use of m-health applications that include gamification elements and perceived health behaviour changes. The green trend line, on the other hand, has a 0 gradient which shows that there is no relationship between the use of m-health applications that do not include gamification elements and perceived health behaviour changes. In other words, the inclusion of gamification elements in m-health applications makes it more likely that use of the application will result in positive health behaviour changes.

Using Cramer's V to determine the strength of the correlation between health behaviour changes in students and the use of m-health applications that include gamification elements results in a V of 0.16, which shows that there is correlation, but this correlation is relatively weak.

C. *First Secondary Research Question*

The first secondary research question asked which element of gamification, if any, particularly drives health behaviour changes in students. In order to answer this question it must be determined whether there is correlation between the type of gamification elements included in used m-health applications and perceived health behaviour changes in students. The collected data outlining the relationship between the level of agreement that health behaviours changed and the type of gamification included in the m-health application can be seen in Table 5, while the expected values should there be no correlation can be seen in Table 6.

Table 5: Collected Responses of Level of Agreement that Health Behaviours Changed vs the Type of Gamification Elements Included

	A/B	P	L	LB	VR	N/S	Other	TOTAL
Strongly Disagree	0	0	0	0	1	0	0	1
Disagree	1	2	1	1	1	0	0	6
Neutral	3	1	1	1	1	1	0	8
Agree	3	3	2	0	0	0	1	9
Strongly Agree	2	2	1	2	0	0	0	7
TOTAL	9	8	5	4	3	1	1	31

Table 6: Expected Responses of Level of Agreement that Health Behaviours Changed vs the Type of Gamification Elements Included Under the Assumption That There is No Correlation

	A/B	P	L	LB	VR	N/S	Other	TOTAL
Strongly Disagree	0.29	0.26	0.16	0.13	0.10	0.03	0.03	1
Disagree	1.74	1.55	0.97	0.77	0.58	0.19	0.19	6
Neutral	2.32	2.06	1.29	1.03	0.77	0.26	0.26	8
Agree	2.61	2.32	1.45	1.16	0.87	0.29	0.29	9
Strongly Agree	2.03	1.81	1.13	0.90	0.68	0.23	0.23	7
TOTAL	9	8	5	4	3	1	1	31

In Tables 5 and 6, A/B stands for Achievements/Badges, P stands for Points System, L stands for Levelling-Up System, LB stands for Leaderboards, VR stands for Virtual Rewards, and N/S stands for Narrative/Story. From the collected responses in Table 5 and the expected responses in Table 6, the chi-squared test results in a p-value of 0.73. In other words, there is a 73% chance that the collected data or data more extreme could occur simply by chance. The data is therefore not statistically significant and no statistical conclusions about the entire sample can be made based on it.

The collected data in Table 5 can be displayed graphically, as seen in Figure 2. Figure 2 shows the number of responses in each level of agreement that health behaviours changed per each of the different types of gamification elements. Trend lines have been added for each type of gamification element. These trend lines represent the observed relationship between the types of gamification elements included in the used m-health applications and how strongly the respondents perceived their health behaviours changed. The majority of the trend lines have a positive gradient, which indicates that there is a positive relationship between their inclusion in m-health applications and health behaviour changes in students. Achievements/Badges has the largest gradient, indicating that it has the largest impact on health behaviour changes in students when it is included in m-health applications. The Narrative/Story trend line has a gradient of 0, which means that there is no relationship between its inclusion in m-health applications and health behaviour changes in students. Interestingly, the Virtual Rewards trend line has a negative gradient which indicates that there is a negative relationship between its inclusion in m-health applications and health behaviour changes in students.

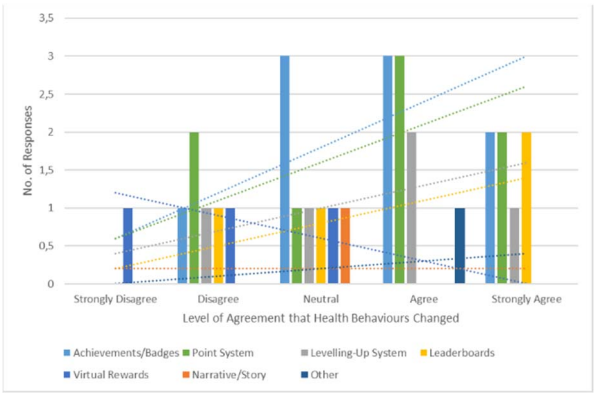


Figure 2: Levels of Agreement that Health Behaviours Changed by Types of Gamification Elements

Using Cramer's V to test the strength of the correlation between the type of gamification elements included in m-health applications and health behaviour changes in students shown in Figure 8, returns a V value of 0.08. This value represents an extremely weak, essentially negligible, correlation between the type of gamification elements

included in m-health applications and health behaviour changes in students.

D. Second Secondary Research Question

The second secondary research question asked what types of m-health applications are being used by students and what types would Non-Users like to use. In order to answer this question the types of m-health applications being used by Users and the types of m-health applications Non-Users would be open to using must be examined to determine if there are types that are more significant than others.

Table 7: Collected Response of Types of Used m-Health Applications vs Types Non-Users Would Use

	Users	Non-Users	TOTAL
Fitness	16	11	27
Diet	8	8	16
Sleep	3	12	15
Medication Management	1	3	4
Negative Habit Cessation	0	4	4
Other	1	0	1
TOTAL	29	38	67

Table 8: Expected Responses of Types of Used m-Health Applications vs Types Non-Users Would Use Under the Assumption That There is No Difference Between the Two Groups

	Users	Non-Users	TOTAL
Fitness	11.69	15.31	27
Diet	6.93	9.07	16
Sleep	6.49	8.51	15
Medication Management	1.73	2.27	4
Negative Habit Cessation	1.73	2.27	4
Other	0.43	0.57	1
TOTAL	29	38	67

Using the collected responses from Table 7 and the expected responses from Table 8, the chi-squared test results in a p-value of 0.045. Therefore, it can be said that the collected data is statistically significant and statistical conclusions about the entire sample can be made based on it.

In order to better compare the collected responses in Table 7 against the expected responses in Table 8, both sets of data have been displayed graphically in Figure 3. By comparing the collected responses against the expected responses, it is clear that the major differences occur in the Fitness and Sleep categories. There are significantly more Users using and less Non-Users interested in using Fitness applications than expected. There are also substantially less Users using and more Non-Users interested in using Sleep applications than expected.

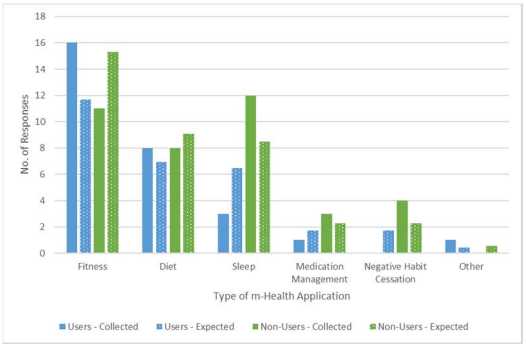


Figure 3: Collected vs Expected Responses of Types of Used m-Health Applications and Types Non-Users Would Be Willing to Use

E. Third Secondary Research Question

The third secondary research question asked how the inclusion of gamification elements in m-health applications affects their adoption rates. In order to answer this question, whether students would be more or less likely to adopt an m-health application if it contains gamification elements must be looked at, as can be seen in Table 9.

Table 9: How Likely Students are to Adopt an m-Health Application if it Contains Gamification Elements

	Users	Non-Users	TOTAL
Less Likely	2	0	2
Neutral	6	2	8
More Likely	12	13	25
TOTAL	20	15	35

Performing a chi squared test and calculating Cramer's V on the collected data in Table 9 results in a p-value of 0.18 and a V of 0.07. Therefore, it can be concluded that there is no correlation between whether a student is a User or Non-User and the change in likelihood that they would use an m-health application if it contained gamification elements. Therefore, the combined total of both the Users and Non-Users can be used to analyse the trend, as can be seen in Figure 4. Figure 4 shows a significant positive relationship between students' likelihood to adopt an m-health application and whether it contains gamification elements.

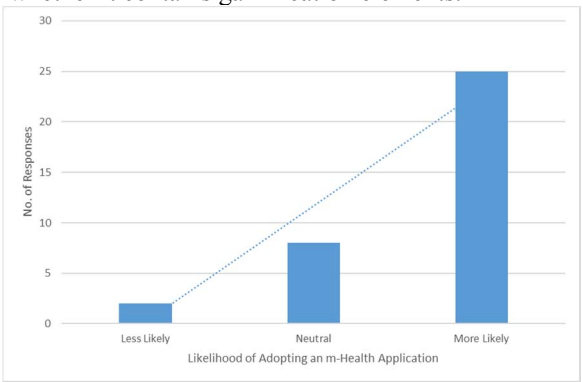


Figure 4: Change in Likelihood of Adopting an m-Health Application if it Contains Gamification Elements

V. Findings & Implications

A. Main Research Question

In answer to the question "how does the inclusion of gamification elements in m-health applications affect health behaviour changes in students?" the collected data appears to represent that the inclusion of gamification elements in m-health application does have a positive, although relatively weak, effect on health behaviour changes in students. However, this conclusion is based on data that is not statistically significant and therefore cannot be considered statistically representative of the entire population. This finding is in line with what was expected based on the facts that gamification shares many overlaps with persuasive computing (Deterding et al., 2011; Fogg, 2003; Llagostera,

2012; Oinas-Kukkonen & Harjumaa, 2009)[7][27][28][29] and gamifications proven effectiveness with the Net Generation (Cronk, 2012; McGrath & Bayerlein, 2013)[4][30]. This finding also backs up the theories of D. King et al. (2013)[31] and Cugelman (2013)[5] that gamification would be an effective tool for changing health behaviours.

A major implication of this key finding is that m-health applications should include gamification elements if they aim to be more successful at changing the health behaviours of students. Designers of m-health applications should be aware of this finding when creating m-health applications so that it can be incorporated into the application. This will make their applications more effective. Students looking to change their health behaviours should also be aware of this key finding as incorporating this information into their decision to use m-health applications could maximise their chances of successfully changing their health behaviours.

B. First Secondary Research Question

In answering the first secondary question "which element of gamification, if any, particularly drives health behaviour changes in students?" the collected data appears to represent that there is likely no gamification element that particularly drives health behaviour changes in students due to the incredibly weak correlation between the two. However, this conclusion is based on data that is not statistically significant and therefore cannot be considered statistically representative of the entire population. Reviewed literature did not make any mention of specific gamification elements that are more effective than others. The likely reason for this is that different gamification elements are applicable in different situations, even within specific categories such as m-health, and thus the finding is line with this.

The implication of this finding further extends that of the findings of the main research question. When m-health application designers are incorporating gamification into their applications, the type of gamification they incorporate will likely not have any effect on the effectiveness of their application, as long as some form of gamification is incorporated. This means that m-health application designers can try to differentiate their applications with the types of gamification that they incorporate without fear of making their applications less effective.

C. Second Secondary Research Question

In answering the question "what particular types of m-health applications are being used by students and what types would non-users like to use?" the collected data appears to represent the fact that Fitness applications are being used most and Non-Users would be most open to using Sleep applications. This conclusion is based on statistically significant data and is therefore statistically representative of the entire population. Fitness applications comprise 38% while Sleep applications comprise <1%. Therefore, there is likely an over-supply of Fitness applications and an under-supply of Sleep applications.

The implication of this finding is that more Sleep related m-health applications need to be developed in order to encourage a greater number of students to begin using m-health applications. Developers of m-health applications should be aware of this when developing new or updating m-health applications, as it may encourage more users to adopt their application.

D. Third Secondary Research Question

In answering the question "how does the inclusion of gamification elements in m-health applications affect their adoption by students?" the collected data appears to represent the fact that students are more likely to adopt an m-health application if it contains gamification elements. This finding was to be expected as the Net Generation are comfortable with the idea of gamification and have actually shown a preference for using it in a multitude of different activities in their lives (Epstein, 2013)[32].

The implication of this finding again goes hand-in-hand with the implication of the findings from the main research question. Not only does including gamification in m-health applications make them more effective at changing the health behaviours of students, it also makes them more appealing to students. Therefore, m-health application designers should also be aware of this when designing their applications as they can use gamification to get more users using their application.

VI. Conclusion

This research examined Gamification in m-Health Applications and its Effects on Health Behavioural Changes in Students. This was done by surveying a sample of students from the University of Cape Town and looking at the relationship between their use of gamification in m-health applications and how they perceived their health behaviours to have changed as a result.

The purpose of this research was to begin to fill in the gap regarding the use of gamification in the health sector, particularly m-health applications, in the hope that it would begin to shed some light on possible solutions to changing individuals' health behaviours. The key finding of this research, that the inclusion of gamification elements in m-health applications has a positive effect on health behaviour changes in students, has fulfilled this purpose. The key finding shows that gamification may indeed provide a possible solution to changing individuals', especially students, negative health behaviours.

Other findings of this research showed that the type of gamification element included in an m-health application does not affect its ability to change health behaviours in students; there is a demand amongst Non-Users of m-health applications for Sleep related applications while the most popular type of m-health application is Fitness; and students are more likely to use an m-health application if it contains gamification elements.

However, there were limitations of this research, especially with regards to the small sample size. This resulted in findings that may not necessarily be representative of the entire population, and thus a recommendation for future research would be to ensure a larger sample size in order to produce more statistically significant findings. Other

recommendations for future research include determining why gamification has this positive effect on health behaviour changes in students and determining why different gamification elements appear to not have different levels of health behaviour change associated with them.

Developers of m-health applications should be aware of the findings of this research and should take them into account when designing m-health applications in order to make them more effective at changing health behaviours of students and to increase their adoption by students. Students looking to change their health behaviours should also be aware of the key finding of this research, that gamification in m-health applications has a positive effect on health behaviour changes in students, in order to make better informed decisions and maximise their chances of improving their health behaviours.

REFERENCES

[1] Mercer, S. W., & Tessier, S. (2001). A qualitative study of general practitioners' and practice nurses' attitudes to obesity management in primary care. Health Bulletin, (59), 248–253.

[2] Martin, T. (2012). Assessing mHealth: Opportunities and Barriers to Patient Engagement. Journal of Health Care for the Poor and Underserved, 23(3), 935–941.

[3] Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamification work? - A literature review of empirical studies on gamification. In Proceedings of the Annual Hawaii International Conference on System Sciences (pp. 3025–3034).

[4] Cronk, M. (2012). Using gamification to increase student engagement and participation in class discussion. In World Conference on Educational Multimedia, Hypermedia and Telecommunications (pp. 311–315).

[5] Cugelman, B. (2013). Gamification: What it is and why it matters to digital health behavior change developers. Journal of Medical Internet Research, 15(12).

[6] King, D., Greaves, F., Exeter, C., & Darzi, A. (2013). "Gamification": influencing health behaviours with games. Journal of the Royal Society of Medicine, 106(3), 76–78.

[7] Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: defining "gamification." In Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments (pp. 9–15).

[8] Domínguez, A., Saenz-De-Navarrete, J., De-Marcos, L., Fernández-Sanz, L., Pagés, C., & Martínez-Herráiz, J. J. (2013). Gamifying learning experiences: Practical implications and outcomes. Computers & Education, 63, 380–392.

[9] Groh, F. (2012). Gamification: State of the Art Definition and Utilization. In Proceedings of the 4th Seminar on Research Trends in Media Informatics (pp. 39–46).

[10] Nicholson, S. (2012). A User-Centered Theoretical Framework for Meaningful Gamification. Games+L.

[11] Simões, J., Redondo, R. D., & Vilas, A. F. (2013). A social gamification framework for a K-6 learning platform. Computers in Human Behavior, 29(2), 345–353.

[12] Huotari, K., & Hamari, J. (2012). Defining Gamification - A Service Marketing Perspective. In Proceedings of the 16th International Academic MindTrek Conference: Envisioning Future Media Environments (pp. 17–22).

[13] Liu, C., Zhu, Q., Holroyd, K. a., & Seng, E. K. (2011). Status and trends of mobile-health applications for iOS devices: A developer's perspective. Journal of Systems and Software, 84(11), 2022–2033.

[14] Free, C., Phillips, G., Watson, L., Galli, L., Felix, L., Edwards, P., … Haines, A. (2013). The Effectiveness of Mobile-Health Technologies to Improve Health Care Service Delivery Processes: A Systematic Review and Meta-Analysis. PLoS Medicine, 10(1).

[15] Wu, A. C., Carpenter, J. F., & Himes, B. E. (2015). Mobile health applications for asthma. The Journal of Allergy and Clinical Immunology: In Practice.

[16] Free, C., Phillips, G., Felix, L., Galli, L., Patel, V., & Edwards, P. (2010). The effectiveness of M-health technologies for improving health and health services: a systematic review protocol. BMC Research Notes, 3(1), 250.

[17] Jones, C., Ramanau, R., Cross, S., & Healing, G. (2010). Net generation or Digital Natives: Is there a distinct new generation entering university? Computers and Education, 54(3), 722–732.

[18] Leung, L. (2004). Net-Generation Attributes and Seductive Properties and Internet Addiction. CyberPsychology & Behavior, 7(3), 333–348.

[19] Oblinger, D., Oblinger, J., & Lippincott, J. (2005). Educating the Net Generation. Brockport Bookshelf.

[20] Sandars, J., & Morrison, C. (2007). What is the Net Generation? The challenge for future medical education. Medical Teacher, 29(2-3), 85–88.

[21] Berdichevsky, D., & Neuenschwander, E. (1999). Toward an ethics of persuasive technology. Communications of the ACM, 42(5), 51–58.

[22] Davis, J. (2009). Design methods for ethical persuasive computing. In Proceedings of the 4th International Conference on Persuasive Technology (p. 6).

[23] Fogg, B. J. (1998). Persuasive computers: perspectives and research directions. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 225–232). ACM Press/Addison-Wesley Publishing Co.

[24] Fogg, B. J. (1999). Persuasive technologie301398. Communications of the ACM, 42(5), 26–29.

[25] Lockton, D., Harrison, D., & Stanton, N. (2008). Design with intent: Persuasive technology in a wider context. In Persuasuive Technology (pp. 274–278).

[26] Saunders, M., Lewis, P., & Thornhill, A. (2009). Research Methods for Business Students. Pearson Education Limited.

[27] Oinas-Kukkonen, H., & Harjumaa, M. (2009). Persuasive systems design: Key issues, process model, and system features. Communications of the Association for Information Systems, 24(1), 485–500.

[28] Fogg, B. J. (2003). How to Motivate & Persuade Users. In CHI 2003: New Horizons.

[29] Llagostera, E. (2012). On Gamification and Persuasion. In Proceedings of SBGames 2012 (pp. 12–21).

[30] McGrath, N., & Bayerlein, L. (2013). Engaging online students through the gamification of learning materials: The present and the future. In Australasian Society for Computers in Learning in Tertiary education (ASCILITE) (pp. 573–577).

[31] King, D., Greaves, F., Exeter, C., & Darzi, A. (2013). "Gamification": influencing health behaviours with games. Journal of the Royal Society of Medicine, 106(3), 76–78.

[32] Epstein, Z. (2013). Enterprise Gamification for Employee Engagement. Retrieved April 28, 2015, from http://www.marioherger.at/attachments/137_EnterpriseGamificationPaper-ZoeEpstein-IthacaCollege.pdf

IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.

Determining Key Factors that Lead to the Adoption of Password Managers

Raymond Maclean and Jacques Ophoff
Department of Information Systems
University of Cape Town
Cape Town, South Africa
mclray002@myuct.ac.za, jacques.ophoff@uct.ac.za

***Abstract*—Passwords form part of our daily routine and even though there are alternative authentication mechanisms, such as biometrics, passwords stubbornly persist. Passwords have been around for the last few decades, but also the various problems associated with users trying to create passwords that are strong and secure. Users are faced with a cognitive burden in managing passwords which often leads to poor password practices or users recycling passwords across various accounts. While there is no anticipated end to the use of passwords, scholars have identified that passwords need to be better supported – one such method is using a password manager. There is a wealth of technical research relating to password managers, which has led to drastic improvements and the maturing of the technology. However, there is a little research on why people would choose to adopt password managers. To explore these factors, this research uses an adapted version of the Unified Theory of Acceptance and Use of Technology (UTAUT2) that includes trust as an additional construct. Using empirical data, the results of the study show that performance expectancy, habit, and trust are key factors in the intention to adopt a password manager.**

Keywords—technology adoption, password manager, information security, trust, UTAUT2, PLS-SEM

I. Introduction

Passwords are used daily by almost every person to access a multitude of accounts, systems and websites. Initially, people only had a few passwords to remember, but with the growth of technology, the sheer volume of accounts with corresponding passwords has increased to the point where keeping track of each password is a burden. With cyber-crime on the rise, the requirements for creating a secure password for each system further complicates the problem, especially as people are reusing or recycling passwords.

While there have been attempts to replace passwords and alternative authentication methods, passwords stubbornly persist. While there is no single solution to the problem, one recommendation is to better support the use of passwords. One such method is using password managers, while they have been around for quite some time and are recommended by security experts, there still seems to be little uptake to using a password manager. While there is research into the various types of password managers, proposals, security concerns and recommendations, there is very little research on why some people adopt password managers. This paper intends to determine the key factors; therefore this paper will not address any technical security concerns of the tools.

The remainder of this paper is organised as follows. First, the problem behind passwords will be presented, this section includes password security, defining password managers, outlining the various available types and related work in the field. Next, the theoretical framework is discussed, this includes the hypotheses and conceptual model. The following chapter will address the research methodology that was used. This is followed by a section on data analysis and a discussion of the results. Lastly, the conclusion summarises the research contributions.

II. Background

In the following subsections, some context around passwords, password problems and an outline of password managers is presented.

A. Password paradigm

Passwords are an integral part of every person's life; they are used daily to access a plethora of online services, systems, devices and computers. As these services have exponentially expanded over the last few decades, the number of login credentials and passwords that users need to recall has drastically increased. Although the death of passwords has been predicted by various key figures, security managers and corporate companies over the last two decades [1]–[4], passwords persist and will most likely remain for quite some time. The persistence of passwords has been acknowledged and irrespective of the ongoing attempt to replace passwords with a worldwide longing to have them replaced, they remain part of our daily lives [5]. It has been argued that no single solution or "silver bullet" would be the answer to the problem, but rather that, a "best-fit" solution would need to be adopted and as passwords would endure for the foreseeable future [5]. A recent report from Cybersecurity Ventures envisions that the "total universe of passwords will likely grow from approximately 90 billion today to 300 billion by 2020" [6, p. 2]. With an understanding of the password epitome, the next section will address password security.

B. Password security

Research into password security related problems dates back to 1979 [7]; this has allowed scholars to contribute a wealth of research in the field over the last few decades. Researchers conducted a systematic literature review that found there has not been a paradigm change in password management for over thirty-five years [8]. The security of passwords remains a significant problem with varying requirements for creating secure passwords, such as password length, alphanumeric characters, special characters and the use of passphrases.

Research into the "characteristics of over 6 million passwords" specifically looked into "password length, password composition, and password selection" [9, p. 130].

Further research into the composition of passwords set out by authentication designers indicated that passwords should not contain any username details, advising that they need to be several characters long and consist of uppercase, lowercase digits and special characters [10]. Gray, Franqueira and Yu [11] found four factors that influenced the recollection of secure passwords, further adding to the dilemma of password security. The requirements for complex passwords and the need to recall passwords places an increase in the cognitive demand of users to have secure passwords for each login. The quality of passwords is exceptionally lacking, with users often recycling passwords and struggling to recall their passwords [12]. More recent research found that people would re-use both complex and repeated passwords at a rate of "1.7" to "3.4" passwords across a spectrum of websites [13, p. 175]. Considering the password paradigm outlined in the previous section, and the challenges faced with password security, there is motivation "to better support the use of passwords" [5, p. 8], one method being the use of password managers.

C. *Password managers defined*

Password managers are "programs used to generate, encrypt, and store passwords for a client-side user" [14, p. 18]. Password managers make use of a master password to unlock a database of more complex passwords, decreasing the "cognitive burden" of users [15], [16]. It is accepted that "Password managers remove the effort from password management" [17, p. 1]. A password manager starts with the user and a master password. The master password unlocks the password manager system, allowing the user to access the secure database and functions of the password manager. Password managers add the benefit of only needing to recall a single secure master password, while the more complex or system generated passwords are stored securely. The password manager can then interact with the various login pages to either auto login or pass the account details to the required system.

D. *Types of password managers*

Password managers now encompass a comprehensive range of password manager schemes across a broad platform of devices, operating systems and technologies that cover client-side programs and mobile apps to cloud-based solutions. Password managers are available as open-source or closed-source packages. Three categories of password managers have been clarified: "desktop manager, online manager and portable manager" [18, p. 234], but there are some password managers that were provided by vendors of browsers, third parties and network-based "where passwords are backed up to the cloud and synced across the user's devices" [19, p. 449]. There is a wide selection of password managers, some noteworthy mentions from previous research [14] include: Encryptr, Passbolt and LastPass for online and cloud-based password managers. For mobile devices, there is a broad selection of mobile apps such as 1Password, Dashlane, KeePassMobile, iCloud Keychain, LastPass, mSecure, OpenIntents Safe for Android, PadLock and Roboform2Go. Client-side password managers include HandyPassword, KeePass2, Padlock, Password Safe and RoboForm to name a few. There is also the availability of browser-based plugins: Password Maker, Password Multiplier and PwdHash. For script-based password managers, there is the option of Password Composer while Password Generator is a considered as a bookmarklet based password manager.

E. *Related work*

Alkaldi and Renaud [20] researched the adoption and rejection of smartphone security tools in 2016; the tools included screen locking functionality, anti-malware applications and password managers, they concluded that smartphone users were not using the available security tools. The authors wanted to "model security behaviours in order to understand adoption or rejection of these tools", their adopted model showed "a number of important factors informing smartphone security intentions" but needed further work to "validate the model with Smartphone owners"[17, p. 142].

Later research then focused on the adoption and rejection of smartphone password managers using "reviews from application stores representing the opinions of users who chose to trial password managers" [17, p. 2]. Various factors that impacted adoption and rejection were found through an online survey with 352 respondents about "password manager use and exploring factors that encourage or discourage password manager adoption" [17, p. 3].

More recent research in 2017 investigated user's considerations in the adoption of password managers through an online survey that encompassed 248 paid participants [21]. The newer research also focused on the examination of forty-five emotions felt by users when using password managers "since emotion has been identified by work in psychology and communications as influential in other risk-laden decision-making" [21, p. 1]. The results of the study found that "convenience" and "usefulness" were part of the main factors leading to the adoption of password managers while security concerns were cited by users that did not use the tool [21, p. 1]. The authors also noted that the "purpose of such tools is often misunderstood by both "users" and "non-users" [21, p. 15]. The analysis of the emotions indicated that users of password managers were "likely to feel secure, admiring and energetic, and less likely to feel suspicious when using their password manager to log into a website" [21, p. 15].

III. THEORETICAL FRAMEWORK

There are a broad variety of models and theories that cover individual acceptance of technology. The Unified Theory of Acceptance and Use of Technology (UTAUT) was created when researchers empirically compared eight user acceptance models in early 2003 [22]. "UTAUT has served as a baseline model and has been applied to the study of a variety of technologies in both organizational and non-organizational settings" [23, p. 158]. The model has key constructs that are linked to use behaviour. The original UTAUT model was further extended in 2012 to a second-generation model named UTAUT2 to address the consumer acceptance and use of Information Technology [23].

A revised UTUAT model based on "trust and acceptance of cloud computing" was used in which the author determined that "trust establishment was the main barriers to adopt cloud services and applications" [24, p. 133]. It can be argued that trust of cloud-based password managers would also influence the adoption of password managers. Trust was a clear underlying theme for participants in research

conducted by Alkaldi and Renaud [17]. Trust is supported by Karole, Saxena and Christin [18] who found that users need a certain level of trust in third-parties when using an online password manager, while users were more likely to trust portable password managers given that they used on their own local devices which they had control over. In remote password storage "there is considerable trust in the third party since it holds all user passwords" [25, p. 320]. Trust of third-parties has been questioned, with the assumption that "the third-party cloud provider can be trusted" [26, p. 314], while more recent research conducted on the adoption of password managers by experts in computer security, found that trust also seemed to be an issue [27].

The researcher proposes an adapted version of the UTAUT2 model with "Trust" as an additional construct with a direct impact on "Behavioural Intention".

A. Hypotheses development and conceptual model

The research model will use an adapted version of the UTAUT2 model with the constructs that are outlined in the subsections below, for this study the various factors that moderate the relationship of each underlying construct will not be tested.

1) Performance Expectancy

Performance expectancy is "the degree to which an individual believes that using the system will help him or her to attain gains in job performance" [22, p. 447]. Performance expectancy affects behavioural intention.

- H1: Performance expectancy has a positive impact on the intention to adopt password managers.

2) Effort Expectancy

"Effort expectancy is defined as the degree of ease associated with the use of the system" [22]. Effect expectancy affects behavioural intention.

- H2: Effort expectancy has a positive effect on the intention to adopt password managers.

3) Social Influence

Social Influence is described as "the degree to which an individual perceives that important others believe he or she should use the new system" and the underlying construct has the "explicit or implicit notion that the individual's behavior is influenced by the way in which they believe others will view them as a result of having used the technology" [22, p. 451]. Social influence affects behavioural intention.

- H3: Social influence has a positive effect on the intention to adopt password managers.

4) Facilitating Conditions

"Facilitating conditions is the degree to which an individual believes that an organizational and technical infrastructure exists to support use of the system" [22, p. 453]. Facilitating conditions includes "aspects of the technological and/or organizational environment that are designed to remove barriers to use" [22, p. 453]. Facilitating conditions affects behavioural intention,

- H4: Facilitating conditions has a positive impact on the intention to adopt password managers.

5) Hedonic Motivation

Hedonic Motivation is defined as the "fun or pleasure derived from using a technology, and it has been shown to play an important role in determining technology acceptance and use" [23, p. 161]. It has a direct effect on behaviour intention.

- H5: Hedonic motivation has a positive impact on the intention to adopt password managers.

6) Price Value

Price Value can be described as the effect that "cost" and "pricing structure" influences the "consumers" use of technology were they comprehend the advantages against the cost of the technology [23, p. 161]. Price value affects behavioural intention.

- H6: Price value has a positive impact on the intention to adopt password managers.

7) Habit

"Habit has been defined as the extent to which people tend to perform behaviors automatically because of learning" [23, p. 162]. Habit affects behavioural intention.

- H7: Habit has a positive impact on the intention to adopt password managers.

8) Trust

Trust is defined as "the belief that you can trust someone or something" and that there is a perceived level that "something is safe and reliable" [28]. Trust was proposed as an additional construct in an extended UTAUT model where it was observed that it had a direct effect on behaviour intention and that "trust establishment was the main barriers to adopt cloud services and applications" [24, p. 133].

- H8: Trust has a positive impact on the intention to adopt password managers.

Based on the above discussion the conceptual research model predicts several factors which influence the adoption of password managers.

IV. Methodology

Given the limited time constraint for the honours research project and the rate at which technology trends change, the timeframe of this research project will be cross-sectional, it will review data on password manager adoption factors using existing literature, along with an online survey to gather data from respondents on their adoption factors as it currently exists.

The target audience consisted of random participants in various fields in technology or IT companies, the main set of respondents predominantly consisted of a large set of students, along with a small subset of staff at a large South African university. Due to the risk of the possible low usage of password managers, the questionnaire also targeted non-users by gathering information about their perceived option on password managers and its anticipated use.

The questionnaire consisted of 31 questions that would take approximately five to ten minutes to complete. To ensure research validity and reliability, the wording for each question was based on the UTAUT/UTAUT2 questions [22], [23] and the work on the extended UTAUT model with the trust construct [24]. The questions were measured using a 7-

point Likert scale. Demographic questions were then asked, followed by questions on account usage and online behaviour based on the work of Fagan, Albayram, Khan and Buck [21]. The survey was published online using the Qualtrics platform. The questions were checked, and a dry run was conducted to check for any errors. The survey link was distributed through official mailing lists within the university, while external participants were emailed and asked to distribute the link to the survey.

V. Data analysis and results

A total of 265 responses were recorded in Qualtrics over a two-week period, 3 participants were under the age of 18 and had to be removed from the survey. 71 participants did not complete the survey, and the incomplete responses were removed from the dataset, leaving a total of 191 responses with data that was used for analysis.

A. Demographic information

The demographic information provided by the 191 participants that completed the survey included gender, age, level of education and level of computer proficiency. Most of the responses (52.88%) were male, closely followed by female respondents (43.46%), while 3.65% of participants preferred not to answer. Most of the participants (53.40%) are between the age of 18 to 25 years old, the second highest respondents (26.70%) were between 35 to 54 years old while there was a small group (18.32%) between the age of 26 to 34. Only 1.57% of participants were between 55 to 65 years, and there were no participants over the age of 65.

Most participants are very well educated with no participants having less than high school education, 32.46% had a 4-year college degree, while 26.53% of the participants had some college diploma. 23.56% of the participants followed closely with high school / General Educational Development (GED) while the remainder of respondents (10.99%) had a master's degree, 7.33% indicated a 2-year college degree, while 3.14% had a doctoral degree, only one respondent (0.52%) held a professional or medical degree.

Most of the participants (38.74%) are highly proficient in the use of computers with 27.75% being very highly skilled and 23.56% being above average. Only 9.42% of participants considered themselves as average users, and one respondent (0.52%) was recorded as being below average. A summary of the demographic information is provided in Table 1.

Table 1. Demographic data

Demographic	Metric	Percentage	Count
Gender	Male	52.88%	101
	Female	43.46%	83
	Prefer not to answer	3.66%	7
Age	18-25 years old	53.40%	102
	26-34 years old	18.32%	35
	35-54 years old	26.70%	51
	55-65 years old	1.57%	3
	65 years or older	0.00%	0
Level of education	Less than High School	0.00%	0
	High School / GED	23.56%	45
	Some College	21.99%	42
	2-year College Degree	7.33%	14
	4-year College Degree	32.46%	62
	Master's Degree	10.99%	21
	Doctoral Degree	3.14%	6
	Professional / Medical Degree (JD, MD)	0.52%	1
Level of computer proficiency	Very Low	0.00%	0
	Low	0.00%	0
	Below average	0.52%	1
	Average	9.42%	18
	Above average	23.56%	45
	High	38.74%	74
	Very high	27.75%	53

B. Accounts and online behaviour

Most participants spend a considerable amount on time online. 98.43% are online more than five times a week, 1.05% went online about four to five times a week, while only one respondent (0.52%) went online two to three times a week.

The participants were asked if they were ever aware of having an account hacked or compromised. 29.84% of the responses indicated that they were aware of an account being compromised, while 54.97% were not aware of being compromised or hacked and 15.18% were unsure. There is a probability that some participants may not be willing to admit that their accounts were compromised and opted not to answer truthfully.

The number of accounts for internet website or services was grouped into six categories. Most participants either have ten to twenty or fewer accounts (58.64%) while 12.04% had five or fewer accounts. 16.23% of participants have fifty or fewer accounts while only 24 respondents (12.57%) had more than fifty accounts. One respondent (0.52%), indicated that they had no accounts, this may be an error in the response, or a misunderstanding of the question, given that the survey was sent out via email to all participants, indicating that they should at least have access to one account.

For the account usage of participants, on an average week, 41.36% used five or fewer accounts, while 36.65% used ten and 17.80% of participants used twenty or fewer accounts per week. There was an extremely low number of respondents (2.62%) who used fifty or fewer accounts per week. Only one participant (0.52%) used the accounts more than fifty times a week while on the contrast, two participants (1.05%) used none of the accounts.

The survey showed that many of the participants did not have unique passwords or sometimes re-used the same password. 40.84% of the participants had several unique passwords, but sometimes reused the same password, 39.79% had few unique passwords and did not vary them across accounts while only 12 respondents (6.28%) had fifty or less unique passwords. Only 10.47% had a unique password for each account, and 2.62% of participants had one password that they used across each account. Password complexity is not considered, hence even though only 20 respondents used a different password for each account; they may still be low entropy passwords and easy to guess.

C. Data analysis tool

The researcher is using Partial Least Squares Structural Equation Modelling (PLS-SEM) for the research model. Partial Least Squares (PLS) is an alternative analysis method for Structural Equation Modelling (SEM) that is "particularly suited to situations in which constructs are measured by a

very large number of indicators and where maximum likelihood covariance-based SEM tools reach their limit" [29, p. 283]. PLS-SEM is an algorithm that is often used in Information System (IS) research for measuring the relationship of constructs in model-based research using latent variables [30]. The researcher used a tool named SmartPLS (version 3.2.7) for the data analysis. SmartPLS is frequently used in model-based research to "estimate the path coefficients, which calculates the strength of the relationships between independent and dependent variables" [31, p. 62]. "Model estimation delivers empirical measures of the relationships between the indicators and the constructs (measurement models), as well as between the constructs (structural model)" [32, p. 131].

D. Model analysis

The data was imported into SmartPLS and the latent variables Performance Expectancy (PE), Effort Expectancy (EE) Social Influence (SI), Facilitating Conditions (FC), Hedonic Motivation (HM), Price Value (PV), Habit (H), Trust (T) and Behavioural Intention (BI) were added to the model. Before proceeding with the analysis of the results, the algorithm must be checked for convergence. The PLS algorithm was calculated using a maximum of three hundred iterations, the stop criterion changes were assessed in the interim results and showed that the algorithm converged in five iterations, well below the set threshold.

The calculation of the PLS algorithm for the initial model included the path coefficients for the inner model and the outer weights/loadings for the outer model; the results show that most the outer loadings are above the required threshold of 0.7 [32]. The outer loadings of the three indicators were below the 0.7 thresholds and subsequently removed before further analysis. Indicators should only be detached if the values of Composite Reliability and Average Variance Extracted (AVE) are amplified [32]. The indicators were removed, and the PLS algorithm was recalculated, the removal of one indicator increased the AVE of FC from 0.541 to 0.693, while the removal of the other two indicators increased the AVE of PV from 0.477 to 0.809. The three indicators were left out of the model for further reliability and validity testing.

Composite Reliability (CR) is used in SEM-PLS to measure the internal consistency reliability. Hair Jr et al [32] indicate that "this measure of reliability takes into account the different outer loadings of the indicator variables", the authors further advise that "reliability varies between 0 and 1" and that higher values will show "higher levels of reliability" [32, p. 136]. Values of 0.60 to 0.70 are tolerable, while values below 0.60 show "a lack of internal consistency reliability" [32, p. 137]. The CR for the model shows that all constructs pass the composite reliability check with values well above the 0.7 thresholds.

Convergent Validity (CV) is defined as "the closeness with which a measure relates to (or converges on) the construct that it is purported to measure" [33, p. 59]. One method of establishing the closeness of the measurements on the construct is using the AVE [32]. An AVE value of 0.50 or higher is desirable as it would typically allow the construct to explain "more than half of the variance of its indicator", while values less than 0.50 would likely indicate that "more variance remains in the error of the items than in the variance explained by the construct [32, p. 138]. The AVE extracted for the model shows that all constructs are above the 0.50 threshold.

Discriminant Validity (DV) is defined as "the degree to which the measures of different constructs differ from one another" [30, p. 19]. DV is often measured together with CV if constructs are linked [33]. DV can be measured using cross-loadings to check the correlation of the indicators outer loadings or using the Fornell-Larcker criterion that compares "the square root of the AVE values with the latent variable correlations" [32, p. 139]. The performance of both methods was recently studied and found not to be entirely dependable in detecting problems with discriminant validity, as an alternative, Heterotrait-Monotrait ratio (HTMT) was nominated as a more accurate technique [32].

An HTMT report was generated for the model, a correlation close to 1 indicates a lack of DV, the acceptable threshold values for HTMT are 0.90 or if "constructs in the path model are conceptually more distinct" a more "conservative threshold value of 0.85" is proposed [32, p. 141]. H and BI loads at 0.877, below the threshold of 0.90, but very close, this could indicate a possible lack of discriminant validity. To truly assess the loading, the confidence interval of the HTMT can be obtained through a procedure known as bootstrapping. A bootstrap calculation was run on the model, the results of the calculation show the path coefficient for H and BI returning a value of 0.608 with a 97.50% level of confidence, indicating that the two constructs are empirically distant.

E. Hypothesis testing

The p-value is an indicator used by scholars to evaluate significance levels; it designates the likelihood of "erroneously rejecting a true null hypothesis" [32, p. 206]. If a researcher is accepting a significance level of 5%, the desired p-value must be lower than a value if 0.05 for the relationship to be regarded as significant at a 5% level, while for more rigours research scholars adopt a significance level of 1% which then requires a p-value of less than 0.01 to designate that the relationship is important [32]. The model was tested with a complete bootstrapping calculation using 5000 samples [32] to test the hypotheses. The results indicate that there are three hypotheses that are significant at a level of 1%, namely H1, H7 and H8, whereas the other hypotheses such as H2, H3, H4, H5 and H6 are not supported. Table 2 provides an overview of the findings.

Table 2. Overview of findings.

Hypothesis		Path Coefficient	t-Value	p-Value	Significance level	Outcome
H1	PE -> BI	0.326	5.430	0.000	p < .001	Supported
H2	EE -> BI	-0.019	0.354	0.723	-	Not supported
H3	SI -> BI	0.009	0.202	0.840	-	Not supported
H4	FC -> BI	-0.021	0.428	0.668	-	Not supported
H5	HM -> BI	-0.042	0.883	0.377	-	Not supported
H6	PV -> BI	0.056	1.271	0.204	-	Not supported
H7	H -> BI	0.517	10.409	0.000	p < .001	Supported
H8	T -> BI	0.162	3.736	0.000	p < .001	Supported

F. Summary of findings

The most substantial result of the significance test was that "Performance Expectancy", "Habit" and "Trust" have a positive impact on "Behavioural Intention" and the adoption of password managers. Trust strongly supports the additional construct that was proposed in the revised UTAUT model

[24] in section III. The significance test also indicated that "Effort Expectancy", "Social Influence", "Facilitating Conditions", "Hedonic Motivation" and "Price Value" were not regarded as having a positive relationship on "Behavioural Intention as originally theorised.

VI. Discussion

Based on the results of this study, three key factors that lead to the adoption of password managers were significantly supported.

A. Performance Expectancy

The participants in this study had a strong link to performance expectancy; finding password managers useful in their daily life while allowing them to accomplish things more quickly. Surprisingly, the use of password managers also increased the productivity amongst the participants. Convenience and usefulness were also identified in a study of user's consideration of password managers use [21]. Performance expectancy is a positive factor in the adoption of password managers.

B. Effort Expectancy

Participants in the study had a clear and understandable interaction with password managers. The data indicated that most of the participants found password managers easy to use or at least, easy to learn how to make use of password managers and become skillful in its intended use. Ease of use, learning and interaction does not seem to be a factor of password manager adoption, most likely since most users of password managers are very computer literate, well-educated and spend a considerable amount of time online.

C. Social Influence

Social influence seems to not play a role in the adoption factors; participants did not consider people that are important to them to influence their behaviour to start using password managers. Only a handful of participants would prefer to use password managers based on the value that they placed in the opinion of people that mattered to them; this was very closely offset by participants that somewhat to strongly disagreed. This shows that peers affecting the social influence of people are not a major driver in this study for the adoption of password managers.

D. Facilitating Conditions

Participants had the necessary resources and knowledge to use password managers; almost all of the respondents indicated that password managers are compatible with the other technologies that they use. There is also a strong indication that they can get help from others if difficulties arise when using password managers. It seems that users do not need any organisational or technical infrastructure to support the use of password managers.

E. Hedonic Motivation

Most of the participants were neutral when asked if they found password managers entertaining, fun or enjoyable. A small number somewhat agreed that password managers were fun, but more participants disagreed, while a higher number disagreed on the entertainment factor. Fun or pleasure in using password managers is not a strong factor in the adoption of the technology.

F. Price Value

The use of free password managers was very favourable amongst the participants; most people did not wish to pay for a password manager; if asked whether they were reasonably priced, most respondents were impartial. More participants, however, felt that password managers were good value for money and at the current price they offered good value. Given that there are many free and open source password managers, the cost and pricing do not seem to affect the adoption of password managers.

G. Habit

While there was a close correlation of the use of password managers becoming a habit for participants, more strongly disagreed. Many felt that they were not addicted to password managers and did not have to use them. The habit of using password managers to generate and store more secure passwords for sensitive accounts was more predominant with expert users [27]. The results of indicate that habit influences the use of password managers and that with more frequent use, habit will automatically become a part of using password managers.

H. Trust

Trust is a strong indicator with most of the participants; they felt that password managers are trustworthy and that they would adopt password managers if good encryption practices were used, especially with regular and secure backups of the password database. The option of an auditing system or environment also had a very positive impact, along with the good reputation of the password manager. The establishment of trust has a positive impact on the adoption of password managers [21].

Performance expectancy, habit and trust influence the use and adoption of password managers. Most participants indicated their intention to continue using password managers into the future and always try to use the technology daily on a more frequent basis.

VII. Conclusion

While there is a wealth of knowledge and prior research on password managers regarding prior shortcomings, various exploits and vulnerabilities, many of the research outcomes have provided insight into improvements and techniques to safeguard the underlying encrypted databases and systems. Research has shown that irrespective of the drive to replace passwords with other authentication methods, passwords stubbornly remain a part of daily life. The password paradigm shows no signs of slowing down, with the number of user accounts and passwords growing exponentially. Literature has indicated that users persist in using poor password practices and that the cognitive load placed on users to create secure passwords for each account led to recycling passwords across accounts.

Password managers have evolved since they were first conceptualised and matured to a point where they are very well suited to allow users to better support the use of passwords, yet little research has been conducted in the field on the adoption of password managers. This study examined

the key factors that lead to the adoption of password managers and used an adapted UTAUT2 model as the theoretical framework. While suited to predict the factors that lead to the adoption and use of technology, the results of the model seem to indicate that only three key constructs had a positive effect on the behavioural intention for adopting password managers.

Performance expectancy was identified as having a positive effect on password manager adoption with data showing a perception that password managers were beneficial and improved efficiency. A second factor was habit and that the continual use of password managers will lead to enforcing the habit of creating more secure and unique passwords for each sensitive account. Trust, linked closely to reputation, was the third and major factor that influenced the intention to adopt password managers, especially if the password manager has good encryption with regular and secure backup options.

ACKNOWLEDGMENT

This work is based on the research supported wholly / in part by the National Research Foundation of South Africa (Grant Numbers 114838).

REFERENCES

[1] M. Kotadia, "Gates predicts death of the password," *Security*, 2004. [Online]. Available: https://www.cnet.com/news/gates-predicts-death-of-the-password/.

[2] IBM, "IBM News room - 2011-12-19 IBM Reveals five innovations that will change our lives within five years - United States," 2011. [Online]. Available: https://www-03.ibm.com/press/us/en/pressrelease/36290.wss. [Accessed: 17-Mar-2018].

[3] D. Terdiman, "Google security exec: 'Passwords are dead,'" *Security*, 2013. [Online]. Available: https://www.cnet.com/news/google-security-exec-passwords-are-dead/.

[4] S. St Louis, "From security to cloud to AI and IoT: Visionaries from Citrix offer predictions for 2018 | Citrix Blogs," 2017. [Online]. Available: https://www.citrix.com/blogs/2017/11/09/from-security-to-cloud-to-ai-and-iot-visionaries-from-citrix-offer-predictions-for-2018/. [Accessed: 17-Mar-2018].

[5] C. Herley and P. Van Oorschot, "A research agenda acknowledging the persistence of passwords," *IEEE Secur. Priv.*, vol. 10, no. 1, pp. 28–36, 2012.

[6] S. Morgan and J. Carson, "The world will need to protect 300 billion passwords by 2020," 2017.

[7] R. Morris and K. Thompson, "Password security: a case history," *Commun. ACM*, vol. 22, no. 11, pp. 594–597, Nov. 1979.

[8] V. Taneski, M. Hericko, and B. Brumen, "Password security - No change in 35 years?," in *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2014 - Proceedings*, 2014, pp. 1360–1365.

[9] C. Shen, T. Yu, H. Xu, G. Yang, and X. Guan, "User practice in password security: An empirical study of real-life passwords in the wild," *Comput. Secur.*, vol. 61, pp. 130–141, 2016.

[10] A. Das, J. Bonneau, M. Caesar, N. Borisov, and X. Wang, "The tangled web of password reuse," in *Proceedings of 2014 Network and Distributed System Security Symposium*, 2014.

[11] J. Gray, V. N. L. Franqueira, and Y. Yu, "Forensically-sound analysis of security risks of using local password managers," in *2016 IEEE 24th International Requirements Engineering Conference Workshops (REW)*, 2016, vol. 5, pp. 114–121.

[12] D. Florencio and C. Herley, "A large-scale study of web password habits," in *Proceedings of the 16th international conference on World Wide Web - WWW '07*, 2007, p. 657.

[13] R. Wash, E. Rader, R. Berman, and Z. Wellmer, "Understanding password choices : How frequently entered passwords are re-used across websites," *Proc. Twelfth Symp. Usable Priv. Secur. (SOUPS 2016)*, no. Soups, pp. 175–188, 2016.

[14] C. Luevanos, J. Elizarraras, K. Hirschi, and J. Yeh, "Analysis on the security and use of password managers," in *2017 18th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT)*, 2017, pp. 17–24.

[15] Z. Li, W. He, D. Akhawe, and D. Song, "The emperor's new password manager: security analysis of web-based password managers," *23rd USENIX Secur. Symp. (USENIX Secur. 14)*, pp. 465–479, 2014.

[16] L. Zhang-Kennedy, S. Chiasson, and P. Van Oorschot, "Revisiting password rules: Facilitating human management of passwords," in *eCrime Researchers Summit, eCrime*, 2016, vol. 2016–June, pp. 81–90.

[17] N. Alkaldi and K. Renaud, "Why do people adopt, or reject, smartphone password managers?," *Eur. Work. Usable Secur.*, p. 15, 2016.

[18] A. Karole, N. Saxena, and N. Christin, "A comparative usability evaluation of traditional password managers," *Int. Conf. Inf. Secur. Cryptol.*, vol. ICISC 2010, pp. 233–251, 2010.

[19] D. Silver, S. Jana, D. Boneh, E. Chen, and C. Jackson, "Password managers: Attacks and defenses," in *Proceedings of the 23rd USENIX Security Symposium*, 2014, pp. 449--464.

[20] N. Alkaldi and K. Renaud, "Why do people adopt, or reject, smartphone security tools?," *Proc. Tenth Int. Symp. Hum. Asp. Inf. Secur. Assur. (HAISA 2016)*, no. Haisa, pp. 135–144, 2016.

[21] M. Fagan, Y. Albayram, M. M. H. Khan, and R. Buck, "An investigation into users' considerations towards using password managers," *Human-centric Comput. Inf. Sci.*, vol. 7, no. 1, pp. 1–20, 2017.

[22] V. Venkatesh, M. G. Morris, G. B. Davis, F. D. Davis, R. H. Smith, and S. M. Walton, "User acceptance of information technology: toward a unified view," *MIS Q.*, vol. 27, no. 3, pp. 425–478, 2003.

[23] V. Venkatesh, J. Y. L. Thong, and X. Xu, "Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology," *MIS Q.*, vol. 36, no. 1, pp. 157–178, 2012.

[24] S. T. Alharbi, "Trust and acceptance of cloud computing: A revised UTAUT model," in *Proceedings - 2014 International Conference on Computational Science and Computational Intelligence, CSCI 2014*, 2014, vol. 2, pp. 131–134.

[25] H. Bojinov, E. Bursztein, X. Boyen, and D. Boneh, "Kamouflage: Loss-resistant password management," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2010, vol. 6345 LNCS, pp. 286–302.

[26] L. Wang, Y. Li, and K. Sun, "Amnesia: A bilateral generative password manager," in *2016 IEEE 36th International Conference on Distributed Computing Systems*, 2016, vol. 2016–Augus, pp. 313–322.

[27] E. Stobert and R. Biddle, "Expert password management," in *Lecture Notes in Computer Science*, 2016, vol. 9551, pp. 3–20.

[28] Cambridge Dictionary, "Meaning of trust in the Cambridge english dictionary," 2018. [Online]. Available: https://dictionary.cambridge.org/dictionary/english/trust. [Accessed: 26-Apr-2018].

[29] M. Haenlein and A. M. Kaplan, "A beginner's guide to partial least squares analysis," *Underst. Stat.*, vol. 3, no. 4, pp. 283–297, 2004.

[30] N. Urbach and F. Ahlemann, "Structural equation modeling in information systems research using partial least square least squares," *Inf. Syst. Res.*, vol. 11, no. 2, pp. 5–40, 2010.

[31] R. Crossler and F. Bélanger, "An extended perspective on individual security behaviors," *ACM SIGMIS Database*, vol. 45, no. 4, pp. 51–71, Nov. 2014.

[32] J. F. Hair Jr, G. T. M. Hult, C. M. Ringle, and M. Sarstedt, *A primer on partial least squares structural equation modeling (PLS-SEM)*. 2017.

[33] A. Bhattacherjee, *Social science research: Principles, methods, and practices*. Scholar Commons, 2012.

Usability of Text-to-Speech Synthesis to Bridge the Digital Divide in South Africa: Language Practitioner Perspectives

Georg I. Schlünz
Human Language Technology Research Group
CSIR Meraka Institute
Pretoria, South Africa
gschlunz@csir.co.za

***Abstract*—We report on two sets of perceptual evaluations of our South African text-to-speech voices by language practitioners. In the first evaluation, we measure baseline quality in terms of how understandable and human-like the voices sound. We also determine baseline usability by asking a series of questions related to accessibility and mainstream application settings. In the second evaluation, we employ the same criteria to compare pronunciation improvements against the baseline. The results indicate success in many areas, but also illuminate room for improvement in others, especially in the cases of the African languages.**

***Index Terms*—Digital divide, accessibility, usability, text-to-speech synthesis, South African languages, perceptual evaluation, language practitioners**

I. Introduction

Text-to-speech (TTS) synthesis, the conversion of electronic text into artificial speech [1], is an important component in assistive technology (AT) to make information and communication more accessible to persons with disabilities. In augmentative and alternative communication (AAC) systems it vocalises messages on behalf of users with little or no functional speech [2]. In screen readers and augmented ebooks it provides an accessible audio narrative of text material for print-disabled users to read [3], [4].

However, the usability of TTS integrated into AT can extend its reach to include not only the disabled community, but also the broader population in a developing country like South Africa that faces multilingual literacy and communication barriers [5]. Riding the wave of the digitalisation of literacy and communication in the 21st century (via web pages, ebooks, apps, etc.), TTS can provide (part of) a solution to reading and learning problems by providing speech as an alternative modality to conventional text. When it is able to cater for the 11 official languages of the country, it becomes an even more empowering tool towards bridging the digital divide.

Essential to the user experience is the ability of TTS voices to synthesise quality speech that is as understandable and natural as possible [1]. The former criterion is formally termed "intelligibility" and measures the correctness with which individual words are pronounced, including word-level (lexical) stress and tone, towards conveying the meaning of the sentence as a whole. The latter criterion, "naturalness", constitutes human-like features such as sentence-level stress (emphasis), intonation (pitch from start to end in phrases), tempo, pausing and breathing.

In previous work [6], we reported on the initial integration of our Speect TTS system [7] (commercialised as "Qfrency TTS" [8]) into an AAC system, screen reader and sample augmented ebooks. We engaged with a small sample of end-users of these applications to determine the usability of the TTS voices in the South African languages. The overall qualitative response was positive, with a pragmatic willingness to adopt the technology, despite certain criteria not yet being met. In the particular case of AAC we were able to quantify the performance more finely using a structured, closed-form survey. Intelligibility scored high for simple test sentences, whereas naturalness were more distributed between the two poles of robotic and human-like. Local language accent was deemed acceptable to good.

In our continued research and development to improve our TTS offering, we aim to increase the number of evaluators for a more representative sample space when doing quality control. We are building feedback mechanisms into the commercial product to obtain input from the market in a more organic fashion, as well as extending networks of collaborators in the accessibility, education, health and publishing sectors to help us out behind the scenes. In the meantime, for this round, we make use of contracted local language practitioners to evaluate our TTS voices. The practitioners can, by virtue of their job description, be viewed as professional custodians of the standards and trends in local language usage. Therefore, we hope that, even in their limited numbers, they can approximate a broad spectrum of the speech and language preferences we can expect to meet with new customers, due to their experience in terms of knowledge and exposure.

This paper relates the perspectives of the language practitioners on the quality of 2 versions of our TTS voices: the baseline and one iteration of pronunciation

TABLE I
TTS VOICE CATALOGUE

Language	Gender	Name	Prompts	Hours
Afrikaans	Female	Maryna	5746	10h58m
Afrikaans	Male	Kobus	4657	07h49m
English	Female	Candice	5251	11h04m
English	Male	Tim	4117	06h43m
Sepedi	Female	Mmapitsi	2609	05h59m
Sepedi	Male	Tshepo	2040	04h41m
Sesotho	Female	Kamohelo	1447	03h25m
Setswana	Female	Lethabo	2337	04h56m
isiXhosa	Female	Zoleka	1705	05h44m
isiXhosa	Male	Vuyo	1580	04h47m
isiZulu	Female	Lindiwe	1708	05h58m
isiZulu	Male	Sifiso	1429	03h30m
isiNdebele	Male	Banele	1454	03h58m
siSwati	Female	Temaswati	1492	05h06m
Tshivenda	Male	Rabelani	2539	04h55m
Xitsonga	Female	Sasekani	0942	01h53m

improvements. Section II describes our voice catalogue, the baseline and the improved versions. Section III is the focus of the paper and expounds on the evaluation methodology and results. Section IV concludes with a summary of findings and recommendations for future work.

II. Implementation

A. Voice Catalogue

We build TTS voices from aligned text and speech training data. The text comprises phonetically-balanced prompts that are read out loud by the voice artists in a professional recording studio. Table I lists the voices in our catalogue with corresponding training data set sizes. They total 16 and cover all 11 official South African languages, with some languages in both genders.

B. Baseline

The Speect/Qfrency TTS system implements statistical parametric speech synthesis. The text frontend performs tokenisation and normalisation [9], grapheme-to-phoneme (G2P) conversion and syllabification, as well as basic prosody prediction using punctuation for phrase breaks and positional and counting features for intonation. The speech backend incorporates the hidden Markov model-based HTS engine [10] to model and synthesise excitation (fundamental frequency), spectrum, voicing strength and duration parameters with a mixed-excitation vocoder.

In particular, the pronunciation lexica of seen words of the baseline TTS voices are based on the Lwazi pronunciation dictionaries [11] that contain roughly 5000 words for each language. The English lexicon is a subset of its roughly 65000-word Lwazi dictionary counterpart. The G2P rules for unseen words are trained on these dictionaries using the Default&Refine algorithm [12].

C. Pronunciation Improvements

In order to increase coverage of seen words, we enlarge the pronunciation lexicon of our English TTS voices using the full 65000-word Lwazi pronunciation dictionary and our Afrikaans TTS voices using the modified RCRL pronunciation dictionary [13], [14] that contains roughly 27000 words. We also extend the lexicons of our African language TTS voices using the NCHLT dictionaries [15], [16] of roughly 15000 words each. This necessitates the standardisation of the phonesets used in the voices, including the creation of phone mappings between the International Phonetic Alphabet (IPA) representations, the Lwazi and NCHLT Speech Assessment Methods Phonetic Alphabet (SAMPA) representations, and old and new internal TTS representations. The resources and methodologies developed in [14], [17] serve as a cross-referencing quality control mechanism.

New G2P rules are trained using the standardised phonesets and updated pronunciation lexica to enable the improved TTS voices to predict the pronunciation of unseen words with greater accuracy.

III. Evaluation

A. Methodology

We perform two sets of evaluations using the contracted language practitioners. The first evaluation requires the practitioners to adjudicate the quality of the baseline TTS voices in an absolute benchmark test. For each language and gender TTS voice, 20 unique phonetically-balanced test sentences are selected from the same domain as the training data and synthesised into audio. Each voice test set is assigned to 5 practitioners, although the same practitioner has to be used in the cases where a language has both a female and a male voice. This bias in the observations is partly mitigated by the unique constitution of each test set. The practitioners must listen to the audio using headphones and cast binary votes (Yes, No) for each of the 20 sentences in an offline spreadsheet, according to certain criteria. The criteria explores aspects of the intelligibility and naturalness of the voices, and are phrased as questions in an absolute sense:

- **Word correct** - Could you hear (distinguish, recognise, identify) most of the words in the sentence? (even if they were not the best quality)
- **Stress/tone** - Are most of the words in the sentence clearly pronounced with the correct lexical stress and/or tone?
- **Sent correct** - Could you understand (follow, interpret) the meaning conveyed by the whole sentence? (even if you had to fill in some gaps)
- **Noise** - Are most of the words in the sentence clearly pronounced with no noisy disturbances like scratches, pops, whistles, etc.?
- **Tempo** - Are most of the words in the sentence pronounced at an acceptable, natural speaking tempo (rhythm, rate)?
- **Intonation** - Does the pitch (intonation) of the voice have an acceptable, natural flow for most of the words from beginning to end in each phrase, as well as most of the phrases from beginning to end in the sentence?

- **Pausing** - Are most of the pauses that are present in the sentence located at appropriate places for an acceptable, natural flow of the speech?
- **Breathing** - Do the breathing sounds of the voice (if any) contribute to an acceptable, natural flow of the speech?
- **Human** - Overall, when the voice speaks this sentence, does it sound more human-like than robotic?

Once the practitioners have finished voting for the sentences, they are required to state their overall conclusions about the voices. These take the form of more binary decisions on the usability of the voices in particular application settings of the technology:

- **First lang** - Overall, is this voice an acceptable representation of a first language speaker?
- **Second lang** - Overall, is this voice an acceptable representation of a second language speaker?
- **News** - If you had to listen to the news (that reports on people, places and events) on a daily basis, instead of reading it silently, would you use this voice to read it out loud for you?
- **Work/School** - If you had to listen to a document at work or school (that contains facts and figures) on a daily basis, instead of reading it silently, would you use this voice to read it out loud for you?
- **Leisure** - If you had to listen to a novel (that dramatises a story for leisure) on a daily basis, instead of reading it silently, would you use this voice to read it out loud for you?
- **No speech** - Imagine you were a person who has no speech (due to a physical impairment or stroke): would this voice be acceptable to use as your communication tool, if it were the only option available? (you type your message in text and the voice speaks it out loud to another person)
- **Blind** - Imagine you were a person who is blind: would this voice be acceptable to use as an alternative reading tool, if it were the only option available? (you use the voice to read your news, documents and novels out loud)
- **Dyslexic** - Imagine you were a person who is dyslexic (you can see written words on a page, but you struggle to take in their meaning): would this voice be acceptable to use as an assistive reading tool, if it were the only option available? (you use the voice to read your news, documents and novels out loud)
- **Illiterate** - Imagine you were a person who can see, but is illiterate: would this voice be acceptable to use as an assistive reading tool, if it were the only option available? (you use the voice to read your news, documents and novels out loud)
- **Learning to read** - Imagine you were a person who can see, but is only starting out in learning to read (like a child in an early grade or an adult at an ABET centre): would this voice be acceptable to use as a learning tool towards reading, if it were the only option available? (you use the voice to read your lesson material, including prescribed books and homework assignments)
- **Preoccupied** - Imagine you were a person who can see and read, but whose visual sense is otherwise preoccupied (you want to read, but you are busy driving to work or exercising at the gym, etc.): would this voice be acceptable to use as an alternative reading tool, if it were the only option available? (you use the voice to read your news, documents and novels out loud)

The second evaluation is a comparative/relative benchmark test between the baseline and the improved TTS voices, in order to confirm that the pronunciation improvements do, in fact, result in better quality voices. We employ the same 20 unique phonetically-balanced test sentences for each language and gender TTS voice and synthesise two samples, one with the baseline and one with the improved voice. This time each test set is only assigned to 3 language practitioners due to cost containment measures and, again, a practitioner is reused over the two genders within a language. The practitioners must listen and select blindly which sample sounds better, or whether they sound the same, though now using a more efficient online web-based survey tool. They must also give reasons for their choice if it is one or the other, according to the same criteria used in the absolute benchmark test, though the questions are rephrased in a comparative/relative sense:

- **Word correct** - It pronounces individual words better
- **Stress/tone** - It has more accurate lexical stress and/or tone
- **Sent correct** - It conveys the meaning of the whole sentence better
- **Noise** - It has fewer noisy disturbances like scratches, pops, whistles, etc.
- **Tempo** - It speaks at a more acceptable, natural tempo (rhythm, rate)
- **Intonation** - It speaks with a more acceptable, natural pitch (intonation) from word to word and phrase to phrase
- **Pausing** - It places pauses in more appropriate, natural places
- **Breathing** - It places breathing sounds in more appropriate, natural places
- **Human** - It sounds more like a human
- **Other** - Other reason

B. Results

We illustrate the absolute benchmark test results for the baseline TTS voices with stacked bar charts in Figures 1, 2, 3 and 4. The counts of votes (horizontal axis) for each criterion (vertical axis) are grouped per Yes (blue), No (red) and N/A (missing) (orange) category. In the analysis that follows, we deem a voice "acceptable" to use when the

number of Yes votes are greater or equal to the number of No votes (#Yes $\geq$ #No).

For the sake of brevity in the exposition, we cluster the Word correct, Stress/tone and Sent correct criteria into "intelligibility" and the Noise, Tempo, Intonation, Pausing, Breathing and Human criteria into "naturalness" where appropriate. The same applies to the use cases. We cluster First lang, Second lang, News, Work/School and Leisure criteria into "mainstream" and the No speech, Blind, Dyslexic, Illiterate, Learning to read and Preoccupied criteria into "accessibility".

The Afrikaans female and male baseline voices in Figure 1 are judged acceptable according to all the intelligibility and naturalness criteria. They are also perceived acceptable in all the accessibility use cases and all the mainstream use cases, except for Work/School in the case of the male voice.

The English female and male baseline voices in Figure 1 are judged acceptable according to all the intelligibility and naturalness criteria, except for Human in both cases. However, these latter scores have a number of N/A (missing) values that could swing the votes otherwise. The voices are perceived acceptable in most accessibility and mainstream use cases. The female voice is not suitable for News, Work/School and Leisure. The male voice is not suitable for News, Work/School and Learning to read.

The Sepedi female and male baseline voices in Figure 2 are judged acceptable according to all the intelligibility and naturalness criteria, except for Pausing in the case of the male voice. Both voices are perceived acceptable in all the accessibility use cases, but only some mainstream use cases. They are not suitable for First lang, News and Leisure. Furthermore, the male voice is not suitable for Work/School.

The Sesotho female baseline voice in Figure 2 is judged acceptable according to all the intelligibility and naturalness criteria. It is perceived acceptable for all the accessibility use cases, but only one mainstream use case. It is not suitable for First lang, News, Work/School and Leisure.

The Setswana female baseline voice in Figure 2 is judged acceptable according to all the intelligibility and naturalness criteria. It is perceived acceptable for all the accessibility use cases, but only some mainstream use cases. It is not suitable for News, Work/School and Leisure.

The isiXhosa female and male baseline voices in Figure 3 are judged acceptable according to two out of the three intelligibility criteria. Neither has adequate Stress/tone. Both voices are judged acceptable according to few naturalness criteria. Neither has adequate Tempo, Pausing and Breathing. The female voice does not have adequate Intonation. Their Human scores are suspect due to the many N/A (missing) values that could swing the votes otherwise. Both voices are perceived acceptable for only some accessibility and mainstream use cases. Neither is suitable for First lang, News, Work/School, Leisure, Dyslexic and Learning to read. Furthermore, the female voice is not suitable for No speech, Blind, Illiterate and Preoccupied.

The isiZulu female and male baseline voices in Figure 3 are judged acceptable according to all the intelligibility and naturalness criteria. They are perceived acceptable in all the accessibility use cases and most mainstream use cases. The female voice is not suitable for First lang, News and Work/School. The male voice is not suitable for Leisure.

The isiNdebele male baseline voice in Figure 4 is judged acceptable according to most intelligibility and naturalness criteria. It does not have adequate Stress/tone nor Breathing. There are a number of N/A (missing) values that could swing the Breathing vote otherwise, but not the Human vote. The voice is perceived acceptable in all the accessibility use cases and all the mainstream use cases, except for First lang.

The siSwati female baseline voice in Figure 4 is judged acceptable according to two out of the three intelligibility criteria. It does not have adequate Stress/tone. The voice is judged acceptable according to only one naturalness criterion. It does not have adequate Tempo, Intonation, Pausing and Breathing, nor is it Human. The latter score has a lot of N/A (missing) values, but they cannot swing the vote otherwise. The voice is perceived acceptable in all the accessibility use cases, except Preoccupied. It is suitable for only one mainstream use case. It is not suitable for First lang, News, Work/School and Leisure.

The Tshivenda male baseline voice in Figure 4 is judged acceptable according to all the intelligibility and naturalness criteria. However, the Human score is suspect due to the many N/A (missing) values that could swing the vote otherwise. The voice is perceived acceptable in all the accessibility use cases, except Learning to read. It is suitable for only one mainstream use case. It is not suitable for First lang, News, Work/School and Leisure.

Finally, the Xitsonga female baseline voice in Figure 4 is judged acceptable according to all the intelligibility and naturalness criteria. It is perceived acceptable in all the accessibility use cases and most mainstream use cases. It is not suitable for First lang and News.

We illustrate the results of the comparative/relative benchmark test between the baseline and the improved TTS voices with stacked bar charts. In Figure 5, the counts of votes (horisontal axis) for each voice (vertical axis) are grouped per votes for the Improved (blue) version, votes for the Baseline (red) version and votes where the 2 versions are perceived Equal (orange). In the analysis that follows, we deem the pronunciation improvements in a voice as "successful" when the number of Improved votes are greater or equal to the number of Baseline votes (#Improved $\geq$ #Baseline).

Figure 6 gives a zoomed out and zoomed in view on the reasons behind preferences towards the Improved versions. Once again, we cluster the Word correct, Stress/tone and Sent correct criteria into "intelligibility" and the Noise,

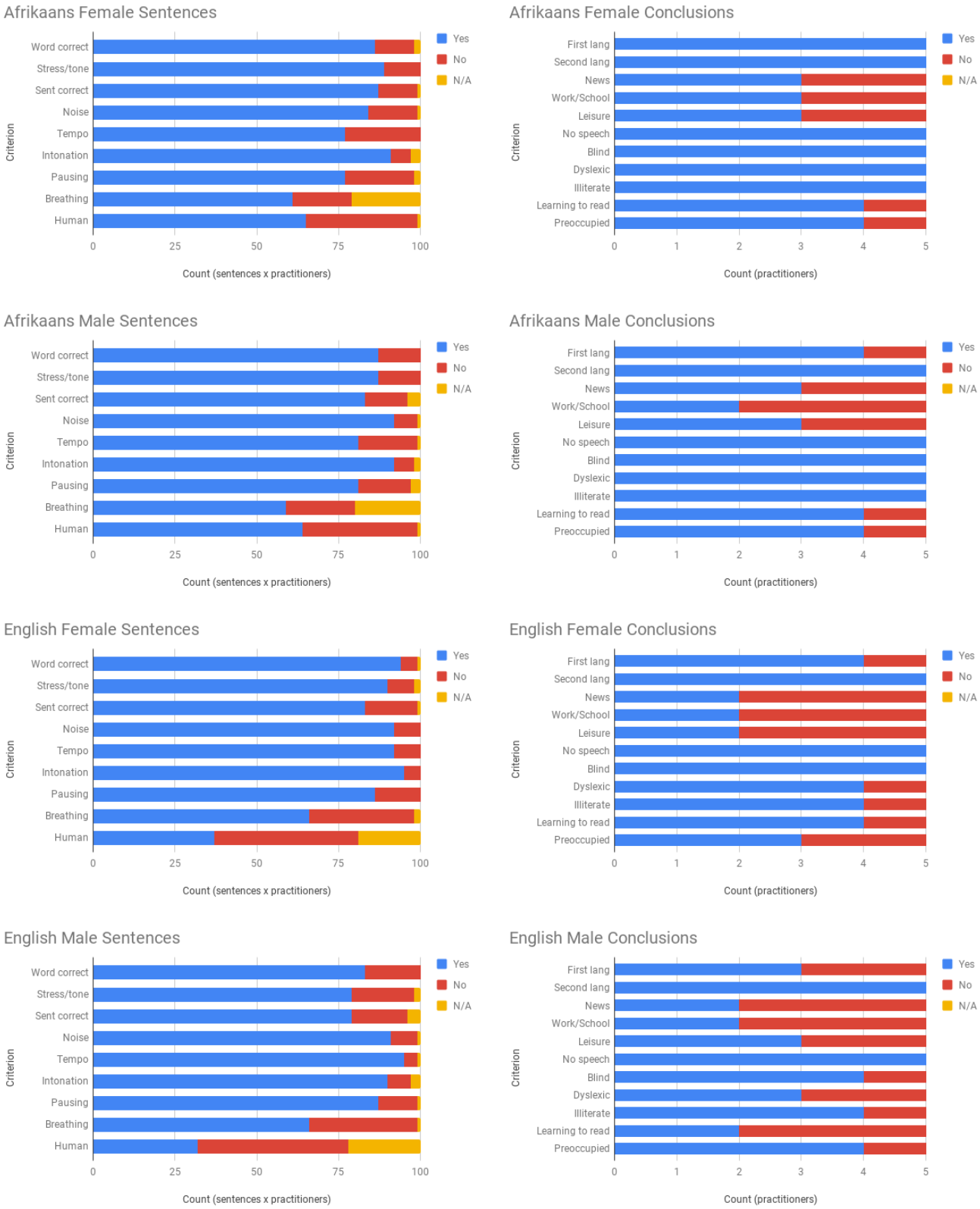


Fig. 1. Afrikaans and English baseline results

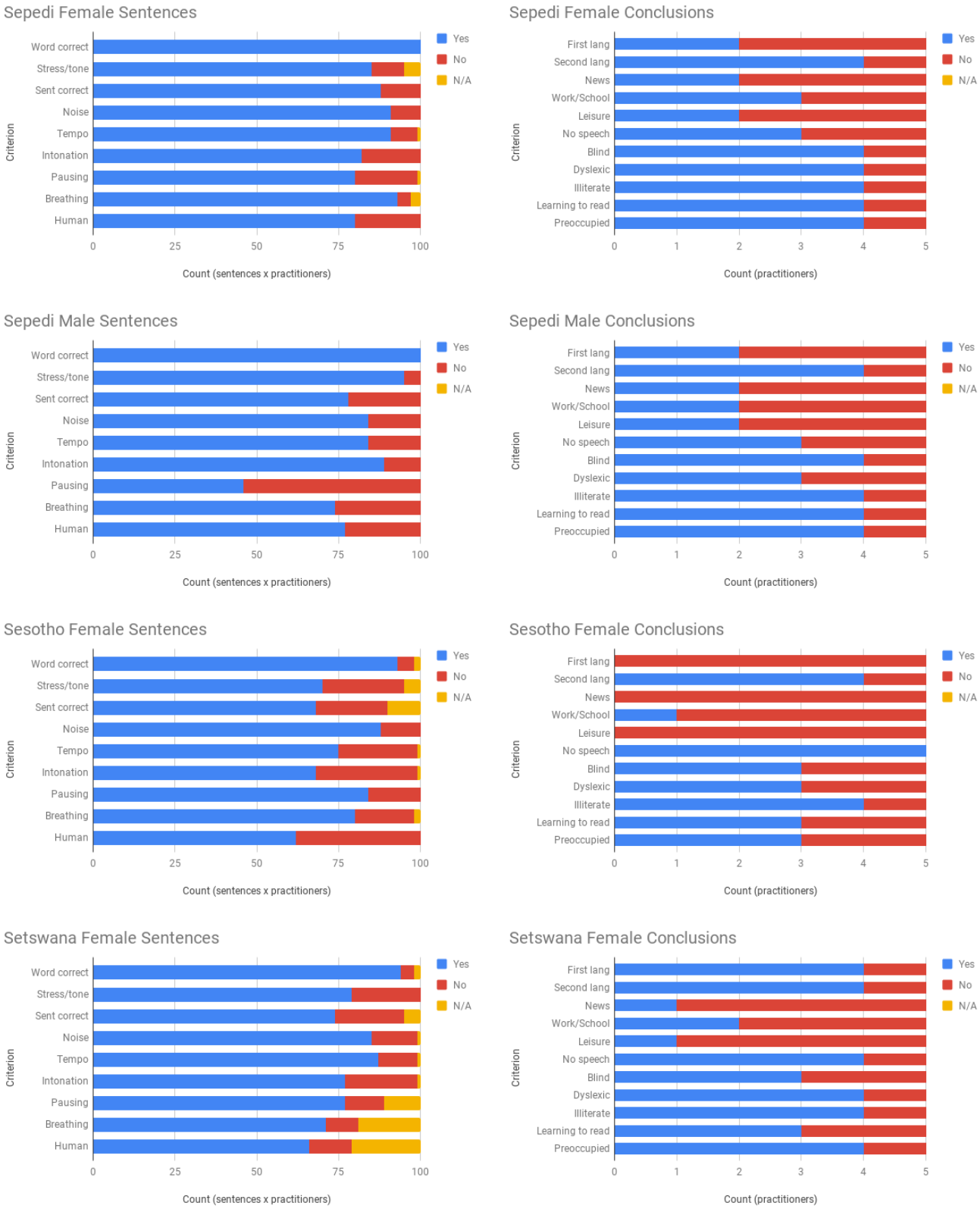


Fig. 2. Sepedi, Sesotho and Setswana baseline results

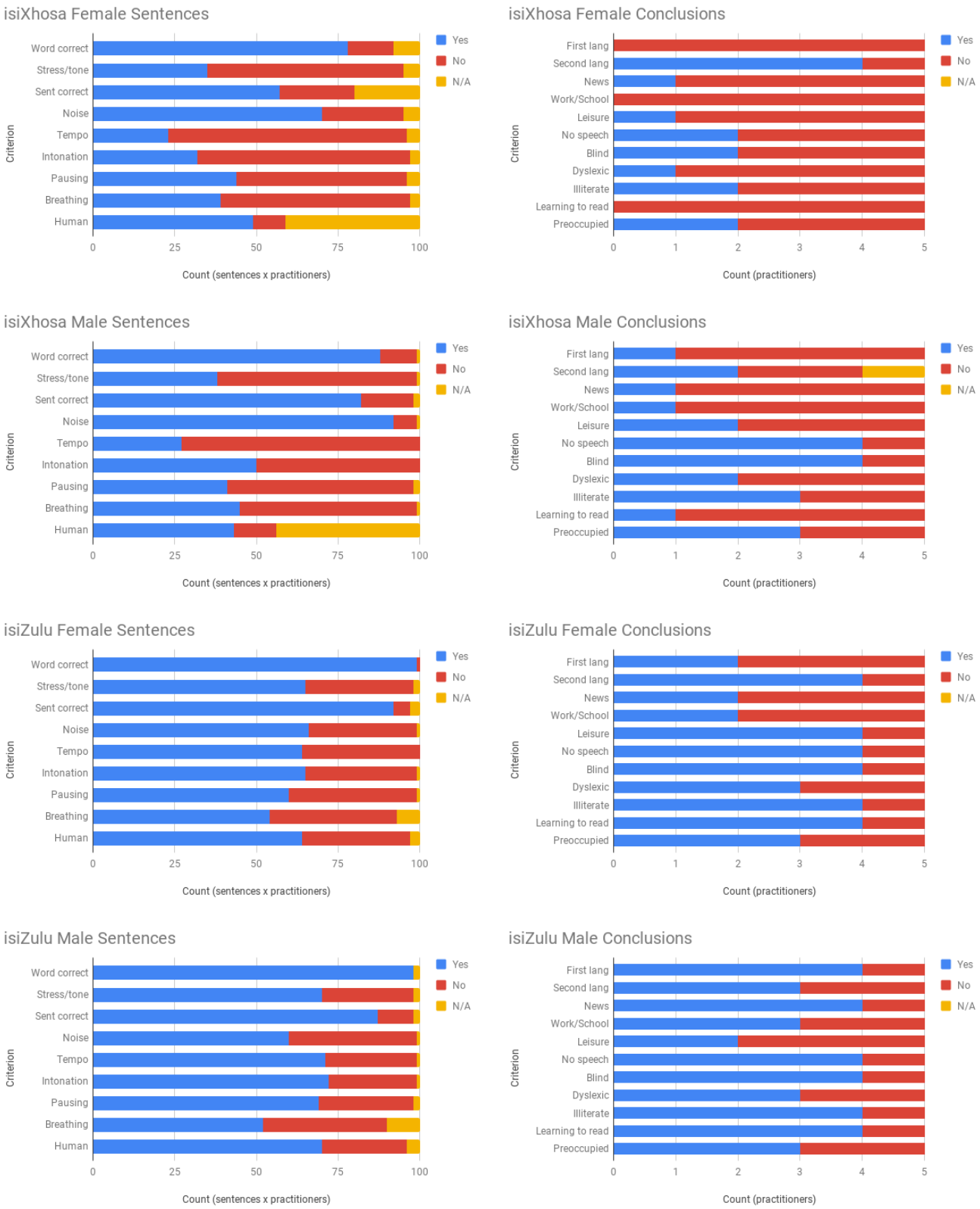


Fig. 3. isiXhosa and isiZulu baseline results

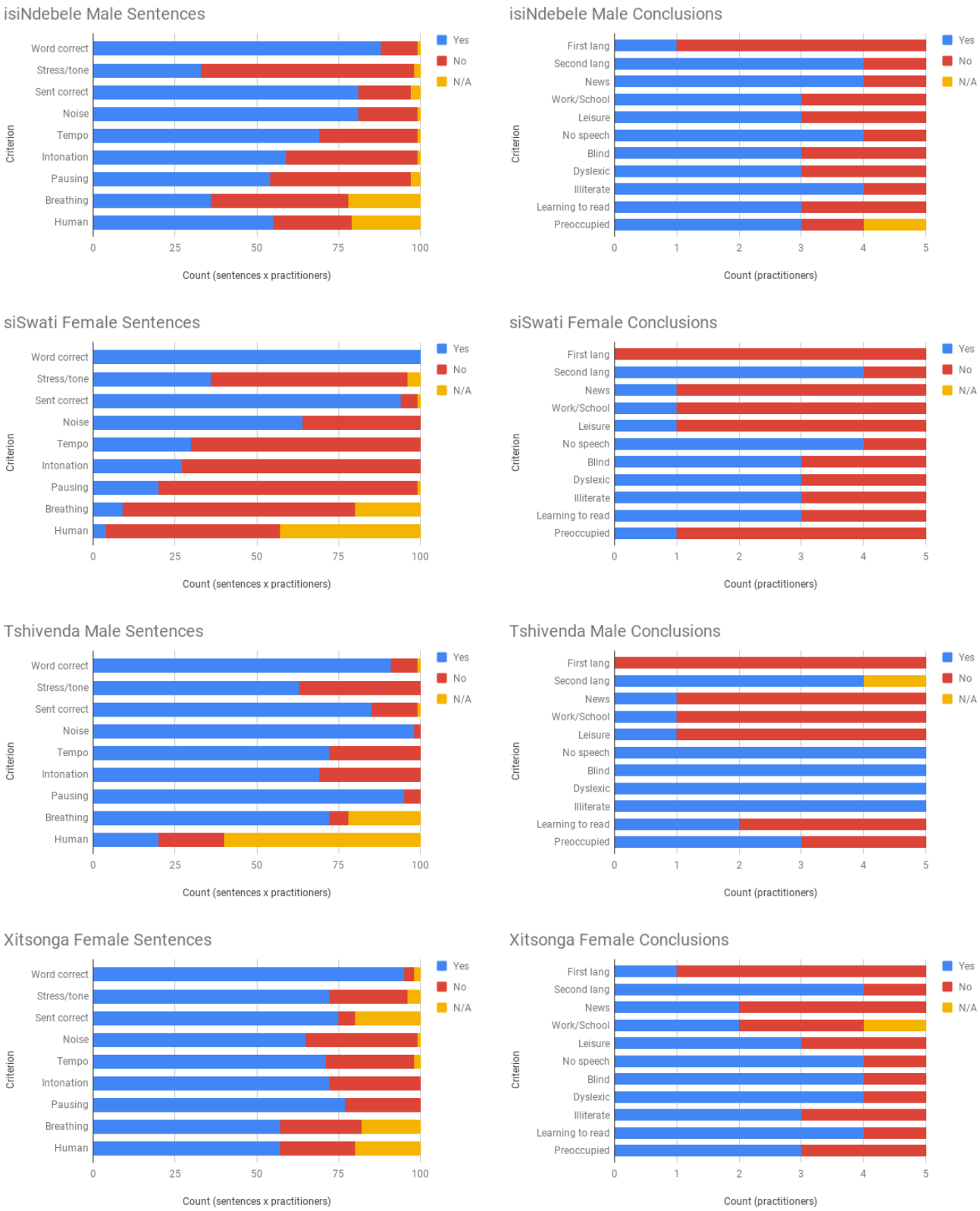


Fig. 4. isiNdebele, siSwati, Tshivenda and Xitsonga baseline results

Tempo, Intonation, Pausing, Breathing and Human criteria into "naturalness" where appropriate.

The pronunciation improvements in the Afrikaans female and male voices are judged successful. The intelligibility reasons given for the preferences towards the Improved versions are Word correct, Stress/tone and to a lesser extent Sent correct. The prominent naturalness reasons are Tempo, Intonation and to a lesser extent Pausing.

The improvements in the English female and male voices are judged successful, although borderline in the case of the male voice. Across both voices, the intelligibility reasons are mostly Word correct and the naturalness reasons Tempo and Pausing. In particular, the female voice also includes Stress/tone and Intonation.

The improvements in the Sepedi female and male voices are judged successful. The seemingly outlying large number of reasons could possibly be explained by the large winning margin of the Improved versions. All the intelligibility and naturalness criteria feature prominently among the reasons.

The improvements in the Sesotho female voice are judged successful, notwithstanding the many Equal votes. The prominent intelligibility reasons are Word correct and Sent correct. The naturalness reasons include all but breathing.

The improvements in the Setswana female voice are judged successful. The major intelligibility reasons are Word correct and Sent correct, while the naturalness counterparts are Noise, Tempo, Intonation and Human.

The pronunciation improvements in the isiXhosa female and male voices are judged successful. The intelligibility and naturalness reasons show similar behaviour to the Sepedi cases in number and criteria coverage.

The improvements in the isiZulu female and male voices are judged successful. The intelligibility reasons are Word correct, Stress/tone and to a lesser extent Sent correct. The prominent naturalness reasons are Tempo, Intonation, Pausing and to a lesser extent Breathing and Human.

The improvements in the isiNdebele male voice are judged successful, despite the many Equal votes. All intelligibility and naturalness reasons feature equally.

The improvements in the siSwati female voice are judged successful, notwithstanding the many Equal votes. Few reasons are given, compared to the isiNdebele case that has a similar winning margin. Only Word correct in intelligibility and Pausing in naturalness are prominent reasons.

The improvements in the Tshivenda male voice are judged successful. The major intelligibility reasons Word correct and Sent correct. The naturalness reasons all feature strongly, except for Intonation.

Finally, the improvements in the Xitsonga female voice are judged successful, although borderline. Relatively, the only prominent intelligibility reason is Sent correct and corresponding naturalness reason is Pausing.

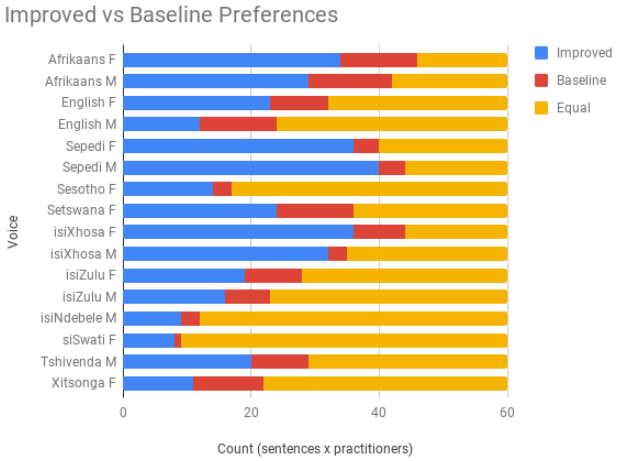


Fig. 5. Improved vs baseline preferences

IV. Conclusion

In the first absolute evaluation, most baseline voices are judged favourably for intelligibility and naturalness. The exceptions are the isiXhosa female and male voices and the siSwati female voice. Most baseline voices are deemed acceptable for use in the accessibility application settings. This confirms the historic market trend of early adoption of TTS in assistive technologies. The exceptions are the isiXhosa female and male voices.

In contrast, the language practitioners are much more conservative in their judgments over use in mainstream application settings. Only the Afrikaans female and male voices, the isiZulu male voice and the isiNdebele male voice have overly positive pronouncements. A particular point of interest is the comparison between the First lang and Second lang use cases. The Germanic language voices score well for both, but the African language voices are not deemed of sufficient quality to pass as first language speakers, only second language ones. This is confirmed by the qualitative feedback from some practitioners. We use the exact same modelling technique for all our voices, hence the reason must lie in the different linguistic structures among the language families. From a theoretical point of view, we are of the opinion that it is the lack of explicit tone modelling that hurts the quality of the tonal African languages. However, overly negative scores for Stress/tone are only reflected for the isiXhosa female and male voices, the isiNdebele male voice and the siSwati female voice.

In the second comparative/relative evaluation, the pronunciation improvements are successful for all the voices, although borderline for the cases of the English male voice and the Xitsonga female voice. The Word correct criterion features consistently as a prominent reason for the preferences towards the Improved versions, with the exception of the Xitsonga female voice. This confirms the theory that larger pronunciation lexica and corresponding

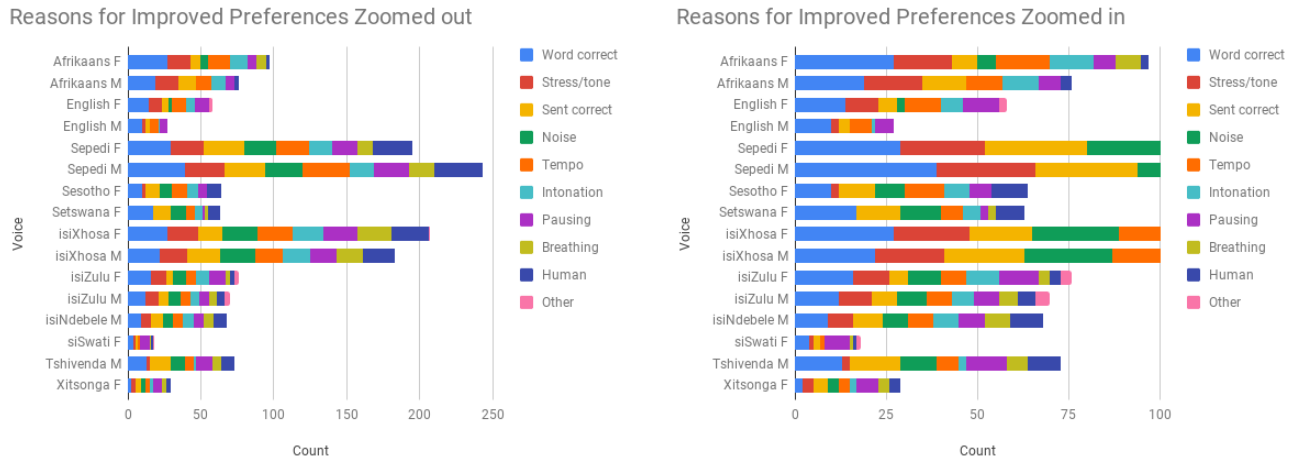


Fig. 6. Reasons for improved preferences

G2P rules should result in better quality pronunciations. Owing to the fact that the phonemes are fundamental units in the implicit modelling technique of our TTS system, the effects of the improved phonemic representations propagate indirectly to other criteria such as Stress/tone, Sent correct, Tempo and Intonation, though they do vary in prominence among the voices.

The pronunciation improvements are being released in our commercial Qfrency TTS offering. The next step on our research and development roadmap is word-level (lexical) stress and tone modelling, in an attempt to bring the quality of the African language voices up to first language standard. There are mainly two approaches to solving the problem: either explicit markup in the pronunciation lexica with an accompanying classifier in the text frontend, or implicit modelling using more advanced deep learning in the speech backend. A combination of the two is also possible. As mentioned in Section I, we also want to expand our sample base for evaluations to allow us to make stronger statistical inferences about our results.

References

[1] P. Taylor, *Text-to-Speech Synthesis*, 1st ed. Cambridge University Press, 2009.

[2] S. Millar and J. Scott, "What is augmentative and alternative communication? an introduction," *Augmentative communication in practice: An introduction*, pp. 3–12, 1998.

[3] L. Rello, H. Saggion, and R. Baeza-Yates, "Keyword highlighting improves comprehension for people with dyslexia," in *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)@ EACL*, 2014, pp. 30–37.

[4] D. Lukeš, "Dyslexia friendly reader: Prototype, designs, and exploratory study," in *Information, Intelligence, Systems and Applications (IISA), 2015 6th International Conference on*. IEEE, 2015, pp. 1–6.

[5] University of Pretoria, "Progress in international reading literacy study (PIRLS) 2016," 2016. [Online]. Available: https://www.up.ac.za/media/shared/164/ZP_Files/pirls-literacy-2016_grade-4_15-dec-2017_low-quality.zp137684.pdf

[6] G. I. Schlünz, I. Wilken, C. Moors, T. Gumede, W. van der Walt, K. Calteaux, K. Tönsing, and K. van Niekerk, "Applications in accessibility of text-to-speech synthesis for South African languages: Initial system integration and user engagement," in *Proceedings of SAICSIT '17*, Thaba Nchu, South Africa, 2017.

[7] J. A. Louw, A. Moodley, and A. Govender, "The Speect text-to-speech entry for the Blizzard Challenge 2016," in *Proceedings of The Blizzard Challenge 2016 Workshop*, 2016.

[8] CSIR, "Qfrency TTS," 2018. [Online]. Available: http://qfrency.com/

[9] G. I. Schlünz, N. Dlamini, A. Tshoane, and S. Ramunyisi, "Text normalisation in text-to-speech synthesis for South African languages: Native number expansion," in *Proceedings of Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, Bloemfontein, South Africa, 2017.

[10] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system version 2.0," in *Proceedings of ISCA SSW6*, 2007, pp. 294–299.

[11] M. H. Davel and O. Martirosian, "Pronunciation dictionary development in resource-scarce environments," in *Proceedings of Interspeech*, Brighton, UK, 2009, pp. 2851–2854.

[12] M. H. Davel and E. Barnard, "Pronunciation prediction with Default&Refine," *Computer Speech & Language*, vol. 22, no. 4, pp. 374–393, Oct. 2008. [Online]. Available: http://dx.doi.org/10.1016/j.csl.2008.01.001

[13] M. H. Davel and F. de Wet, "Verifying pronunciation dictionaries using conflict analysis," in *Proceedings of Interspeech*, Makuhari, Japan, 2010, pp. 1898–1901.

[14] D. R. van Niekerk, "Syllabification for Afrikaans speech synthesis," in *Proceedings of Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, Stellenbosch, South Africa, 2016.

[15] M. H. Davel, W. D. Basson, C. van Heerden, and E. Barnard, "NCHLT dictionaries: Project report," North-West University, Tech. Rep., 2013. [Online]. Available: https://sites.google.com/site/nchltspeechcorpus/home

[16] E. Barnard, M. H. Davel, C. J. van Heerden, F. de Wet, and J. A. C. Badenhorst, "The NCHLT speech corpus of the South African languages," in *Proceedings of the Workshop on Spoken Language Technologies for Under-resourced languages (SLTU)*, St. Petersburg, Russia, 2014, pp. 194–200.

[17] D. R. van Niekerk, C. J. van Heerden, M. H. Davel, N. Kleynhans, O. Kjartansson, M. Jansche, and L. Ha, "Rapid development of TTS corpora for four South African languages," in *Proceedings of Interspeech*, Stockholm, Sweden, 2017.

A Survey on Spectrum Handoff Algorithm for Continuous Connectivity

Duduzile Zamanhlane Masuku
Department of Computer Science
University of Mpumalanga
Nelspruit, Mpumalanga, South Africa
Duduzile.masuku@ump.ac.za

Mthulisi Velempini
Department of Computer
University of Limpopo
Limpopo, South Africa
Mthulisi.velempini@ul.ac.za

Naison Gasela
Department of Computer Science
University of North-West
Mafikeng, North-West, South Africa
Naison.gasela@unw.ac.za

***Abstract*—Cognitive Radio networks (CR) is a new technique that uses available unlicensed spectrum band this is due to the limited number of fixed licensed spectrum bands** [1]**. The main features of the CR are spectrum sensing, management, sharing, and mobility** [2]**. This paper focuses on spectrum mobility for military systems often referred to as spectrum handoff for military systems, which is a process when the CR user changes its frequency of operation due to spectrum occupancy by the licensed user. Spectrum mobility is important in military networks since mission success may depend on reliable communications. In this paper, we propose the use of Fuzzy logic to come up with a decision making handoff scheme. The proposed scheme avoids interruptions caused by the movement of a secondary user and minimizes handoff latency if the SU selects a correct channel. It was observed through the analysis of literature that even though most of the existing schemes can perform handoff successfully, these existing schemes result in slow handoff and are complicated to implement.**

Keywords—Cognitive radio networks, handoff, fuzzy logic

I. INTRODUCTION

Cognitive radio (CR) networks provide high bandwidth to mobile users for heterogeneous wireless networks and provide dynamic spectrum access (DSA) capabilities [3]. Military systems may benefit from DSA but also need waveform selection, protocol selection and a reliable decision-making handoff mechanism [3]. CR networks use the spectrum in a dynamic manner by allowing the radio terminals to operate in the best available frequency band and keep seamless communication requirements while handing off to a better channel to continue with its activities [4].

In CR a spectrum handoff is a process where by a secondary user switches to another spectrum, this could be because of the PU occupying the spectrum at the same time as the SU or an SU may decide to handoff because of some movements by the SU which might cause instability in network connection [5] . During handoff, the main challenge is a slow handoff, which might result in drop in packets and delay transmission of high priority applications [5]. The delay during handoff may result in a compromised quality of services. To avoid all these problems there is a need for an effective decision making handoff scheme that will also implement faster algorithms to avoid signal loss. It is important that the channel selected is able to accommodate the SU for a reasonable amount of time to avoid multiple handoffs; therefore, the handoff algorithm has to be smart enough to avoid interruption by the PU. This may be achieved by spectrum sensing which the handoff algorithm can do. Therefore, the optimal algorithm should be able to maintain QoS, security and provide continuous connectivity for both PU and SU.

When interference is no longer tolerable to PU, or when the QoS required by the SU is not satisfactory, the SU has to handoff the spectrum to quickly change its current channel of operation to occupy an idle channel [5][20]. The fuzzy logic based approach will be able to make an intelligent spectrum handoff decision in a context characterized by heterogeneous information, which will initiate handoff as fast as required by their users. Compared to other capabilities of CR networks such as spectrum sensing, management and spectrum sharing, spectrum mobility has been explored in less-depth by the research community [20].

II. SPECTRUM HANDOFF

In this section, we discuss spectrum handoff in general and the types of handoff.

A change in channels by the SU happens when the PU arrives in channel currently occupied by the SU and whenever there is an interruption to the SU. This often leads to delay to the SU and may cause a drop in packets. Another situation that leads to handoff is the spatial movement of SU, there is a chance of overlap in transmission coverage of the SU with a PU currently using the same channel [6] [7]. An opportunistic user can also cause quality degradation on both the unlicensed user and licensed user when they coexist in the same channel band. In Cognitive networks the rule of thumb is, PUs always has high priority in using the license channel than SU. Spectrum handoff procedures aims to assist SU find suitable free channels to resume the unfinished transmission [7]. There are two types of spectrum handoff based on target cannel selection and are as follows:

A. Reactive Handoff: When a reactive handoff is used, the SU initiates reactive spectrum sensing and reactive handoff mechanism. In a case whereby there is an event that triggers the handoff, the SU will first perform spectrum sensing in other to find a target backup channel [7]. Then communication link is formed with the new available channel. In this case, the targeted channel selection and the handoff action are initiated when there is an event that is requires the SU to change channel of operation. The advantage of this mechanism is that since spectrum sensing is performed on the most appropriate channel, the SU will handoff to the accurate target channel and this might eliminate unnecessary handoffs.

B. Proactive Handoff: In proactive handoff, SU also uses a combination of proactive spectrum sensing and proactive handoff mechanism. Before the handoff occurs, the SU initiates spectrum sensing to see if there is an available channel. The SU uses the knowledge of the PU traffic model to predict a vacant channel before it is interrupted by the PU's arrival. The advantage of this type of handoff is that the handoff delay is very short because of the planning by the SU. Another advantage is that unnecessary handoffs might be avoided [7].

Figure 1 represents the two types of spectrum handoff, which are reactive and proactive handoff.

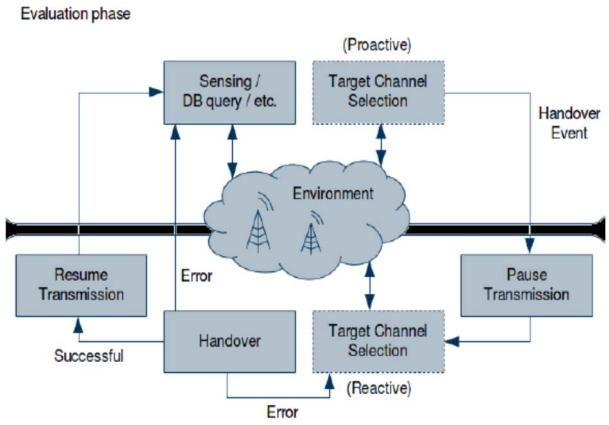


Figure 1: Spectrum handoff types [7]

III. RELATED WORK

In this section, we highlight some of the research efforts that have been made to address spectrum handoff using fuzzy logic. Authors in [8] proposed a spectrum handoff based on mobility, QoS and priority based on the resource usability. The system mostly focuses on mobility of SU. Fuzzy logic and artificial neural networks is used to develop and evaluate the spectrum handoff mechanism [8]. The fuzzy controller was used to estimate values of the parameters fed in to it. The use of artificial neural network helps in getting precision of about 100% in handoff decision. The system is designed for only seven-cell cluster; and it may be extended for more cells cluster.

M. Uddin, et .al [9] used the fuzzy logic concept for optimal handoff decision-making. It gives the clear idea whether handoff is required or not. Artificial Neural Network (ANN) concept is used for the representation of channel and data training in the network for several times to get accurate channel selection. Inputs such as bit error rate (BER) of SU, bit error rate of PU and outputs such as HO (indicates whether handoff is required or not) and P_{SU} decrement (power decrement of SU). Output P_{SU} indicates whether power decrement of SU is required or not. The decision is taken depending on the results given by the rule base system. After the successful decision, SU compares the transmission power levels with PU, if there is a decrease in power levels on the SU then there will be no interference to the PU. Then both the SU and PU can communicate on the same channel or if PU faces interference due to power level of SU, then SU should take handoff.

A spectrum handoff scheme proposed in [9], which is based on approach that works in two ways. In this scheme, the transmission power of SU intelligently adjusted in order to avoid handoff by reducing harmful interference to SU. Then handoff decisions is intelligently implemented using expected holding time (HT) of the channel. This scheme results in reduction of handoff rate as compared to the random selection algorithm [7].

A Feedback Fuzzy Hierarchical Process (FFAHP) was explored to improve decision making for spectrum handoff [10]. A comparison was done in order to assess the performance level of the developed algorithms and the most relevant algorithms in literature. The scheme seeks to use the probability of channel availability, estimated channel time availability, the signal to interference plus noise ratio and bandwidth to make a handoff decision. [10]. The scheme is tested on a real spectrum occupation data taken from the Wi-Fi band and the GSM, which simulated the real behaviour of the PU [8].

Real time and best efforts in high and low traffic levels were used for the evaluation of this scheme. The number of unsuccessful handoff, bandwidth, latency and throughput were considered as the matrices for evaluation [8]. The experiments shows that there is improvement in performance, when using the FFAHP scheme. The FFAHP gives effective and efficient process for selecting frequency channels. Furthermore, the scheme has an advantage of a low number of handoffs, an effective use of the bandwidth, it is faster and result to higher throughput rate these advantages helps in stabilizing the system avoiding unnecessary handoffs [8].

The authors in [11] proposed that a backup channel could be sensed by the SU that will work with the operating channel, which the sensor nodes can sense. The sensor network assisted CR network was proposed as a solution to save energy during handoff. A partial Markov decision process (POMDP) takes into consideration the backup, the current operating channels as well as the remaining energy of the SU in order to perform handoff, the aim is to dynamically decide whether to switch the channel or not in order to increase throughput. Parameters such as detection probability, the maximum energy of the SU, channel idle probabilities of the operating and backup channels are used for simulation [11].

Fuzzy Logic Controllers (FLC) is used in [1] and fuzzy sets such as received power, spectrum load and arrival rate of PU were also used to determine the SU's handoff time. This system is aimed at eliminating unwanted handoffs in LTE-A systems [1]. FLC helps in determining the handoff probability based on the three fuzzy variables. The variables used for evaluation are received power, spectrum load (users) and user's arrival rate of the PU. It is therefore concluded from the results that the higher probability of handoff is a result of the OU arrival and high loading spectrum [1]. When the PU arrives in a channel occupied by the SU, the SU will prefer to change to occupy an available channel where it can continue with its transmission. This system only provides a good solution to avoid the communication interruption caused by the mobility of SU but did not minimise the number of handoffs [1].

A scheme that uses fuzzy logic and combinatorial fusion was used in this schme considered matrics susch as received signal strength, data rate, and network latency for decision-making [12]. This method control strengths of fuzzy logic while maintaining less computational complexity by means of combinatorial fusion analysis that is responsible for handoff decision making [12]. The researchers evaluated the scheme on a three tier networking environment. In this scenario, networks set up for the test are wireless local area networks are over lapping with cellular networks. In the case when there is a need to handoff from the WLANs to cellular, the handoff will occur when connectivity is not sustainable on either network. This handoff scheme uses the signal strength to determine whether it is time to handoff or not [12].

The scheme in [7] was compared with other three schemes for effectiveness. The schemes ware compared on the number of unnecessary handoffs initiated during simulations. During the experiments, it was identified that the Fuzzification and combinatorial fusion scheme resulted in less handoff compared to RSS-base and fuzzy-logic, which initiated 29 handoffs. The scheme initiated the same number of handoffs with the Kalman fuzzy-based scheme [12]. It is therefore evident that the fuzzy schemes are capable of reducing the delay and number of unnecessary handoffs by up to 36% compared to the RSS based handoff schemes [12]. A larger number handoff indicates that a UE is likely to switch between the two networks creating unnecessary handoffs, which may lead to drops of packets and delay [15]. The performance of fuzzification and combinatorial fusion does not show any huge difference as compared to the other vertical handoffs mechanisms. Mechanism compared in [13] still result in a higher number of handoff which leads to lost in connection.

A system that considered only straight-line distances between CPE has and base station is proposed [8]. Change in direction and nonlinear distances were not considered. Minimising the time taken when handing off and minimising the number of handoffs was not considered.

IV. THE PROPOSED SPECTRUM HANDOFF ALGORITHM FOR CONTINOUS CONNECTIVITY

In this section, we highlight fuzzy logic as a tool to be used in designing a decision-making algorithm. This research aims to minimize the rate of spectrum handoff delays, which results from the PU occupying the spectrum and interrupting the SU. We also aim to apply a fuzzy-based optimal spectrum access scheme in military CR network that will help minimize the number of unnecessary handoffs and improve throughput. The handoff will also maintain safety as the SU switches channels. Fuzzy logic is considered as an intelligent tool for making decisions where inputs are uncertain or interpreted qualitatively [14]. Fuzzy logic is also capable of transforming qualitative and heterogeneous information into understandable membership values this makes it easy to create fuzzy rules of the fuzzy inference [14].

Deriving the Algorithm: We will design a handoff initiation algorithm based on fuzzy logic and fuzzy inference systems that can handle imprecisions and uncertainties in data then process multiple parameters for handoff initiation [15]. A fuzzy multiple attributes decision making method and context awareness scheme will be used to design a suitable access network selection function that can handle trade-offs among different handoff metrics including quality of service requirements (such as network conditions, security and system performance), mobile terminal conditions, application requirements and system preference. Power transmitted by the SU will also be used to determine the time to handoff, furthermore is can be reduced to minimise redundant handoffs.

$$SU_{SNR} = 10\log\left(\frac{P_{pu}g(d_{ps})}{\delta^2}\right) \qquad (1)$$

The Shannon-Hartley theorem will be used to determine the power transmitted which will indicate the level of packet transmission [16] (1). This means if QoS is compromised, then an option is to handoff to another channel. A fuzzy logic system (FLS) is able to handle uncertain data and numerical data at the same time in order to produce easily interpreted data this special feature provides uniqueness [17]. Fuzzy set theory and fuzzy logic are able to establish the specifics of the nonlinear mapping [17]. This is done by translating the crisp set theory into to logic that can be then translated to understandable values [18] .

We are therefore, interested in applying fuzzy logic to interpret the transmission into different states, which are followed by different degrees of membership. By using fuzzy logic, a fuzzy-based spectrum handoff scheme that will benefit both the SU and the PU can be designed. Therefore, the aim of this research is to apply a fuzzy-based intelligent spectrum decision making scheme that will allow seamless handoff in the CR network. The fuzzy-based handoff scheme will be suitable for military networks and will consider the spectrum availability (distance of the SU to the PU), application priority and providing quality of services.

Fuzzy logic figure 2 is a good option since it is multi valued-logic and accepts different parameters when making a decision [18]. In general, Fuzzy Logic and Fuzzy decision making consists of three stages such as, Fuzzification, Fuzzy reasoning and Defuzzification [18] .

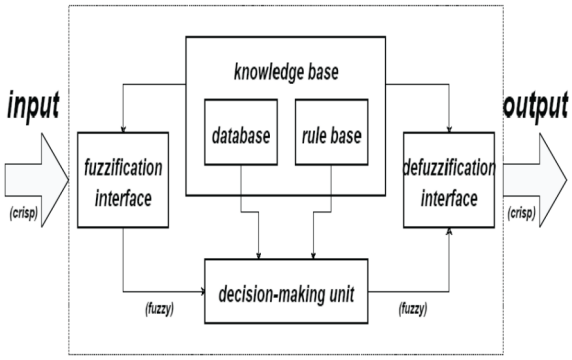


Figure 2: fuzzy logic System [17]

Fuzzy membership function

A membership function is also referred to as a characteristic function and it defines the fuzzy set [19]. While

membership functions for fuzzy sets can be of any type, these functions must satisfy the following constraints;

- MF must range between [0, 1].
- Each $x \in X, \mu A(x)$ should be different.

Commonly, triangular, trapezoidal and Gaussian functions are used as MFs [15]. These parameterized functions are introduced below.

Triangular MF: A triangular MF is a function of a vector, x, and depends on three scalar parameters $a, b, and\ c$ with a<b<c,

The parameters $a\ and\ c$ are considered as x coordinates of the bottom of the triangle and the parameter b represents the x coordinate at the peak of the triangle [20].

Trapezoidal MF: A trapezoidal MF is a function of a vector x, and depends on four scalar parameters $a, b, c,$ and d with, $a < b < c < d,$ as follows:

$$Trapmf\ (x; a, b, c, d) = \begin{cases} 0, & x \le a \\ \frac{(x-a)}{(b-a)}, & a \le x \le b, b \ne a \\ 1, & b \le x \le c \\ \frac{(d-x)}{(d-c)}, & d \le x, \quad d \ne c \end{cases} \qquad (2)$$

In the above equation (2) parameters a and d are the x coordinates at the bottom of the trapezoid and the parameters *b and c* are the x coordinates in fuzzy logic [20].

$$\text{gaussmf}(x; \sigma, c) = \exp(-((x - c)/\sigma)2$$

Gaussian MF is widely used to define fuzzy sets in complicated systems and is a good for modeling non-linear systems [20].

V. ANALYSIS OF LITERATURE

Even though the spectrum handoff in CRs has been studied before, most of the work focused on proactive and reactive handoff. According to literature, a combination of both proactive and reactive has not been explored. The two-way handoff algorithm may help in reducing handoff delays, which may result in increased throughput and higher quality of services for both the SU and PU. According to literature, it is evident that Fuzzy logic can be a useful tool to achieve a reliable decision making algorithm because of the capability of coping with various criteria for decision-making purposes [1]. The weakness of this decision making approach is that it is dependent on the way the membership functions and fuzzy inference rules are set and an additional learning mechanism is needed for effective results.

Even though most decision making algorithms are able to perform handoffs but they still result in number of unnecessary handoffs. A fuzzy logic-based algorithm, Kalman fuzzy-based, Fuzzification, Combinatorial Fusion and RSS based algorithms were compared in [12]. It was observed that the first three approaches reduce handoffs considerably in comparison to the RSS-based, these schemes reduce amount of handoffs for up to 37% and this makes an improves throughput [12]. Which is still not enough since the scheme can still result in slow handoff and are complicated to implement. This may lead to connection loss and drop in packets. Our proposed scheme seeks to overcome the above-mentioned issues.

VI. CONCLUSION

Seamless mobility is important to every network user especially those who are always on the move, as they need best connection at all times, increasing user satisfaction while maintaining a satisfactory level of QoS and seamless connectivity. This is important because of the heterogeneity of characteristics of the wireless networks. It is evident from literature that making decisions for handing off in cognitive radio environment is extremely challenging [21]. In order to reduce delay and loss of data during spectrum handoff a new mobility and connection management scheme needs to be devised [7]. In this research paper, we have proposed a spectrum handoff scheme that will use fuzzy logic to make decision when there is a need to. This is still a work in progress future work includes designing a fuzzy based scheme for spectrum handoff and testing it for effectiveness.

REFERENCES

[1] R. Mardeni, A. Abdulrab, M. Y. A and N. P. Khairil "Efficient handoff spectrum scheme using fuzzy decision making in cognitive radio system," *2017 3rd International Conference on Frontiers of Signal Processing (ICFSP),* 2017.

[2] A. abdulraqeb, R. Mardeni and Y. A. Mohamad, "Analysis of spectrum handoff schemes in cognitive radio network using swarm optimization," in *IEEE 3rd Symposium on Telecommucation Technologies(ISTT)*, 2016.

[3] T. Tang and T. Watson, "Cognitive Radio Networks for Tectical wireless communication," in *DRDC*, Canada, 2014.

[4] V. Mahadevan and J. Zhou, "A Cognitive Radio Based System For Intergrated Ad-HoC Network," in *Computer and Networking Technology*, 2009.

[5] D. Salgado, C. Hernande, V. Molina and F. A. Beltran-Molina, "Intelligent algorithm for spectrum mobility in cognitive wireless networks," *The 7th International Conference on Ambient Systems, Networks and Technologies,* vol. 7, no. 83, pp. 278-283, 2016.

[6] S. Moh, I. Chung and christian, "Spectrum mobility in cognitive radio networks," *IEEE communications Magazine,* 2012.

[7] K. Tiwari and A. Rastogi, "Spectrum Handoff in Cognitive Radio Network," *International Journal of Advanced Research in Computer and Communication Engineering ,* vol. 5, no. 4, pp. 1025-1030, 2016.

[8] S. D. Potdar and K. P. Patil, "Efficient spectrum handoff in CR network based on mobility,QoS and Priority using fuzzy logic and neural networl," in *Conferance on Contemporary Computing*, 2013.

[9] N. A. Lala, M. Uddin and N. A. Sheikh, "Novel Spectrum Handoff in Cognitive Radio Networks Using Fuzzy Logic," *I.J. Information Technology and Computer Science,* pp. 103-110, 2013.

[10] C. A. Hernández-Suárez, L. F. Pedraza-Martínez and E. R. de la Colina, "aprendeenlinea," 2016. [Online]. Available: aprendeenlinea.udea.edu.co/revistas. [Accessed 01 08 2018].

[11] U. Muhammad, S. Muhammad, V.-V. Hiep and I. Koo, "mdpi.com," Open Access Sensors, 2015. [Online]. Available: www.mdpi.com/journal/sensors. [Accessed 26 09 2018].

[12] I. Kustiawan, C.Y. Liu and F. D. Hsu, "Vertical Handoff Decision Using Fuzzification and Combinatorial Fusion," in *IEEE Communication letters*, 2017.

[13] K. Kumar, A. Prakash and R. Tripathi, "Spetrum Handoff scheme with multiple attributes decision making for optimal network selection in cognitive networks," *Digital Communications and Networks,* 2017.

[14] C. A. Hernandez-Suarez, L. F. Pedraza-Martnez and E. R. de la Colina, "Fuzzy feedback algorithm for the spectral handoff in cognitive radio networks," *Revista Facultad de Ingeniería, Universidad de Antioquia,* no. 83, pp. 47-62, 2016.

[15] Y. Nkansah-Gyekye, "A Vertical Handoff Decision Algorithm for Next Generation Wireless Networks," in *International Conference on Broadband Communications Infrmation Technology and Biomedical Applications*, 2008.

[16] R. Mardeni, K. Anuar and H. M. Alias, "Efficient Handover algorithm using fuzzy logic underlay porer sharing for cognitive radio wireless networks," in *IEEE Symposium on Wireless Technology and Application (ISWTA)*, 2013.

[17] B. E. Carvajal-Gamez, H. E. Rubio, V. Meneses and F. Hernandez-Castaneda, "Chapter 26 Speech Recognition Native Module Environment Inherent in Mobile Devices," *Springer Nature,* 2015.

[18] X. Geraldo, "semanticscholar," Federal University of Rio de Janeiro, [Online]. Available: https://pdfs.semanticscholar.org. [Accessed 20 August 2018].

[19] R. Xianyi, Y. Zheng, T. Hu and J. Zhang, "Image Fusion Based on NSCT and Fuzzy Logic," in *International Conference on Multimedia Technology*, 2010.

[20] Y. C. Shin and C. Xu, "books.google," [Online]. Available: https://books.google.co.za/books?isbn=1351835408. [Accessed 21 August 2018].

[21] A. I. Perez-Neira, "Fuzzy-Based Spectrum Handoff in Cognitive Networks," in *International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CrownCom2008)*, 2008.

Interfacing a Brake-by-Wire Simulink Model with SUMO

Erik Martín Campaña, Nils Müllner, Saad Mubeen
Mälardalen University, Västerås, Sweden
ema16003@student.mdh.se, nils.muellner@mdh.se, saad.mubeen@mdh.se

***Abstract*—This paper presents an adapter to link the Matlab/Simulink model of a custom braking system into an open-source road traffic simulator, namely SUMO. The traditional braking models in SUMO work on a coarser level, generally modeling the deceleration of the vehicle based on a single equation. Substituting this model with a detailed system allows adjustments on a very low level, including delays between components, the traction of individual tires, and controllers for the anti-locking system by simulating those in Simulink. Besides introducing the novel adapter linked into SUMO and discussing parameter fitting to mimic the behavior of the traditional braking system, this paper discusses the amount of performance deterioration as the price for simulating a more sophisticated braking system.**

***Keywords*—SUMO, interface, braking simulation, Simulink.**

I. Introduction

Simulation of Urban MObility (SUMO) [1] is an open-source road traffic simulator, capable of microscopic level simulations. It is able to simulate in high detail every vehicle on the road, and it is versatile, capable of modeling various scenarios. Its simulations can contribute to many fields like traffic-flow optimization, pollution reduction and the simulation of autonomous driving [2], [3], [4]. It was first released in 2001 as a simple, open source tool to enable research and continuously improved since then.

Nowadays, a cars' mechanical systems are replaced by functionally equivalent embedded systems. For instance, mechanical brakes are replaced by electronic brakes, also called *brake-by-wire* systems (BBW). A BBW system is a distributed embedded system substituting most pieces of a fully mechanical braking system [5]. While the traditional braking model in SUMO — in the form of a simple equation for the deceleration — mimics the behavior of mechanical brakes, novel BBW systems can be modeled in more detail. For that matter, the current braking model employed by SUMO is unnecessarily unrealistic, relying on one abstract equation for the whole braking process of the vehicle. For instance, i) internal delays within the BBW's internal components, ii) roadslip individual for each tire at each point in time, and iii) control algorithms providing for anti-lock braking system (ABS) functionality can be specified individually for each internal component, for each individual vehicle. As a matter of fact, it could be used in combination with real world testing of new braking systems [6].

A new, accurate model, taking into account all the internal details, allows for more realistic simulations. In this *nanoscopic* level of simulation, any changes can potentially affect the rest of the execution, like a butterfly effect. Since the only parameter that can be defined for the braking system in SUMO is the deceleration, all braking systems are considered to be uniform. In cases where the braking model is relevant for the purposes of research (cf. e.g., [4]), the delays on the braking system are to be accounted for since they could influence the results. This paper introduces a new interface that allows a custom model of the braking system to be utilized. Moreover, the paper discusses the implementation of the newly presented interface.

The range of models that can be connected and interfaced with this interface, may include all the interactions and behaviors of each of the BBW subsystem, including potential delays they have. The resulting interface can be exploited for a novel model-testing method, since models of custom subsystems can be hooked up and tested in this simulation environment.

A. Problem Formulation

The current braking model in SUMO abstracts important information about low-level properties, rendering it unnecessarily unrealistic and leading to inexact simulation results. Developing an interface — allowing for the integration of a more sophisticated and realistic model of the braking system and its internal embedded subsystems — alleviates the abstractions, improving the simulations realism. Equipped with the chance to simulate different models of braking systems, provides the opportunity for a more detailed simulation of the vehicles and enables researchers to perform simulations regarding the braking models themselves or applications that are highly dependent on the braking systems.

B. Research Questions

This paper[1] addresses the following research questions:

1) How can the existing abstract mathematical model for braking in SUMO be replaced by a custom model?
2) What kind of impact does using a custom braking system make in the simulation of SUMO?

C. Approach and Contributions

This paper discusses how a custom braking system can be integrated into SUMO. The tools available are the model of a BBW system provided as a prototype by Volvo cars, and

[1]The work in this paper builds upon the work in [7].

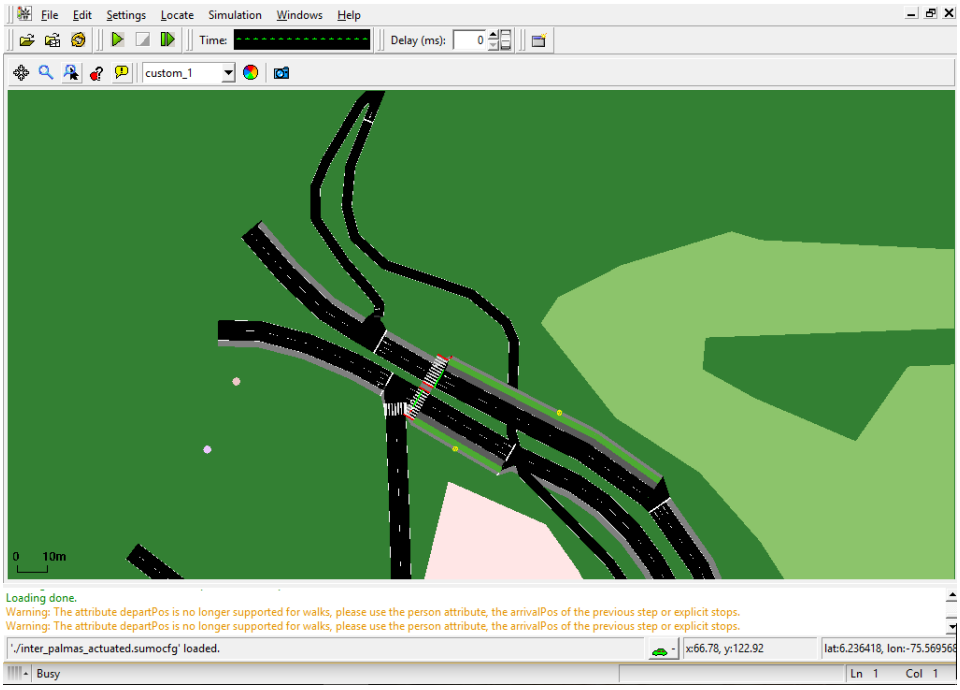


Fig. 1: The SUMO GUI showing a traffic network.

the SUMO simulation environment. The contribution in this paper is the interface/adapter connecting both, integrating the BBW as surrogate for the old braking equation, and fitting the parameters to mimic the behavior of the old system.

Fitting the parameters of the new system such that it surrogates the old braking model demonstrates the impact of this change. The adapter presented is designed to allow for Simulink models in general to be applicable as braking-system surrogates in SUMO. Utilizing the aforementioned prototype model, simulations will be run in order to check the results and adjust the adapter so that the simulation results are as close as possible to the original ones.

D. Outline

The rest of the paper is organized as follows. Section II discusses the details about how SUMO and Simulink work and how they can be employed in order to have a custom braking system simulated in SUMO. Section III discusses the design and rationale behind the adapter that allows to substitute SUMO's braking model by a custom braking model. Section IV shows in detail how the testing of the braking system is accomplished and provides the results. Section V concludes of the paper.

II. Interfacing BBW and SUMO

This section discusses the details of interfacing SUMO and Simulink and how they can be employed in order to have a custom braking system integrated in SUMO.

A. Simulation of Urban MObility (SUMO)

Generally, the traffic modeling is distinguished at four different level, which are described as follows.

1) *Macroscopic level* refers to the traffic flow,
2) *Mesoscopic level* simulations exist between microscopic and macroscopic simulations. Sets of vehicles can be modelled as queues, and vehicles are moved between the queues,
3) *Microscopic level* refers to the behavior of a single vehicle regarding its environment and the driver control [8], and
4) *Nanoscopic level* (or sub-microscopic level) addresses the (sub-)systems within a vehicle. This notation allows to address specific sub-systems on their particular level when perceived as system-of-systems model.

SUMO is a highly portable open source simulation tool, designed to be employed with large road networks. It is able to work at the microscopic level (i.e., simulating every car on the road), mesoscopic level (less detailed but faster) or macroscopic level (highly abstracted but the fastest). This paper focuses on the microscopic level.

SUMO includes a set of tools for different kinds of applications, providing great flexibility. The most important one utilized here allows an online interaction with the simulation, called the *Traffic Control Interface* (TraCI). It allows to modify parameters of the simulation during its execution, which provides for a great amount of possibilities modifying the simulation behavior during run-time. SUMO executes in a console or with a Graphical User Interface (GUI) as shown in the example in Figure 1. The console version performs

naturally faster and is thus selected in this paper. A simulation run is specified in a configuration file before the execution.

B. Simulink

Simulink is a model-based modeling and simulation tool suite that is designed to work seamlessly together with Matlab. It is extensively adopted for modeling and simulating in various fields such as signal processing, power control, robotics, just to name a few. It packs a wide amount of features such as automated code generation, simulation parallelization and custom scheduling. Figure 2 depicts the design interface of Simulink.

The BBW system developed by Volvo and introduced in the MBAT-ARTEMIS project[2] allows for demonstrating the feasibility of the interface. To the best of the authors' knowledge, only once has a Simulink model been interfaced in SUMO [9], [10] which did not concern a model of the braking system. One reason the BBW by Volvo qualifies as suitable candidate is that the system is available as real physical (prototypical) system, as several implementations, as (extended) UML models and as Simulink model, putting the quantitative assessment within a simulation environment like SUMO en par with discrete techniques from software testing (as discussed in the outlook in Section V).

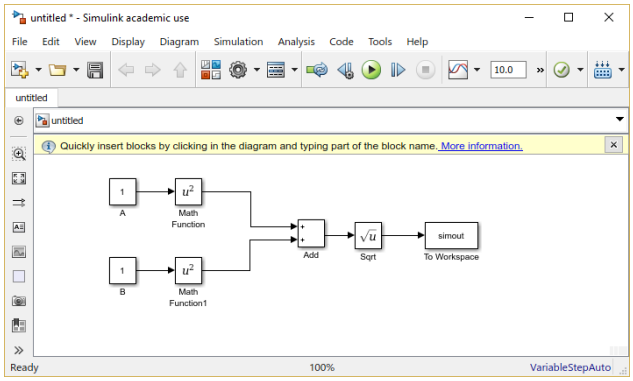


Fig. 2: View of Simulink's workspace.

C. TraCI4Matlab

The selected tool for connecting Simulink and SUMO via TraCI is TraCI4Matlab [11]. It is an Application Programming Interface (API) that allows Matlab to communicate with SUMO's TraCI to interact directly with the executed simulation. Since TraCI works by creating a server where SUMO runs, TraCI4Matlab employs the TCP/IP layer to communicate MATLAB's program that behaves as a client to it.

III. An Adapter Linking the BBW via TraCI to SUMO

The basic structure of the system is shown in Figure 3. The simulation in SUMO and the Simulink model that represents the braking system are connected via the adapter. The three components are explained below.

SUMO simulation ⟷ Adapter ⟷ Simulink model

Fig. 3: An adapter connecting SUMO to the Simulink model.

A. Simulation Environment

Figure 4 shows the selected simulation scenario, consisting of a 500 meters long road with a stop (i.e., initiating braking) at the 450 meters mark. The maximum permitted velocity is 70 m/s and the maximum velocity of the car is 100 m/s. These are arbitrary values selected so that the velocity is not a limiting factor for testing the brakes. Regarding SUMO's behavior when approaching a stop mark, it attempts to commence braking as close to the stop mark as possible before stopping completely. This means, that the car will press and release the brake several times before it stops. After stopping, the car will stay in that position for a set amount of time before starting to move again. In this case, the time that the car is stopped is set to 20 seconds.

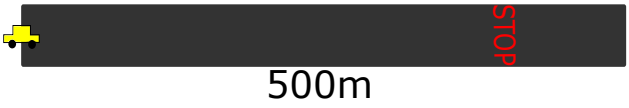


Fig. 4: Test test scenario, driving from left to right.

SUMO has a default seed for random number generation set to 23423. This causes runs to have the same results if other parameters are kept the same. Whenever a variety of scenarios is desired, the simulation is initialized with a different random seed for random number generation every time (e.g., based on the system time-stamp). The command launching the simulation is as follows.

```
sumo -c scenarioPath --remote-port PORT
--random --step-length STEPLENGTH --start
```

The arguments' function is:

a) -c scenarioPath: It loads the configuration file with the extension .sumocfg. This scenario file includes all the information for the simulation like the route file, the networks file or the configuration file.

b) –remote-port PORT: It selects which port is chosen to connect to TraCI. It is only needed when running several instances of SUMO with different simulations at the same time. If it is used, when initializing TraCI in MATLAB, the same port should also be specified.

c) –random: As said before, SUMO's seed for the random number generator is always 23423. If different scenarios are needed, this option should be enabled. It will initialize the random number generator with the system time at that moment.

[2]Combined Model-based Analysis and Testing of Embedded Systems http://www.mbat-artemis.eu/

d) –step-length STEPLENGTH: It sets the length of a SUMO step in seconds. The step length set in all the simulations of this paper is 0.1 seconds. The default step length of 1 second does not offer enough resolution when observing the behavior of the braking system.

e) –start: This argument tells SUMO to initiate the simulation after loading the scenario.

B. Braking Model

This prototype consists of a simple close loop system where for each time step in the Simulink simulation, the step size set in SUMO (0.1 seconds) is multiplied by the same acceleration that was employed in the vehicle description in SUMO (5 m/s^2). This exact prototype model can be seen in Figure 5. The aim of this model is to build the adapter that makes the interaction between SUMO and the braking model possible. It is done this way because this model being very simple and equivalent to the system implemented by SUMO for the braking calculations[3], any difference or problem can be easily tracked. It will also be exploited later to test the impact of the adapter in the simulation.

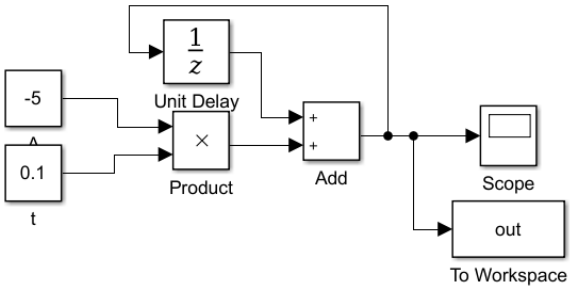


Fig. 5: Simulink version of old system: ($V = v_0 + at$).

C. Design and Implementation of the Adapter

The idea behind this adapter is to substitute SUMO's braking simulation with any braking model that is capable of outputting the speed to MATLAB.

1) Basic operation and implementation: The adapter is the component that intercepts the braking process in SUMO and substitutes it by the one in the model. For that, it follows the following process:

a) Load the SUMO simulation and the braking model in Simulink,
b) Run the simulation in SUMO step by step,
c) When the brake pedal starts being pressed, the Simulink simulation is initialized with the necessary values and run. The results from each step of the simulation are buffered in MATLAB.
d) For each step of the simulation when the brake pedal is pressed, the speed will be set according to the buffered values.
e) If the brake is released, the control is returned to SUMO until the brake pedal is pressed again, in which case it repeats from step c).

All this process is done in MATLAB, which is able to natively communicate with Simulink and via TraCI4Matlab with SUMO.

It is important that the simulation is executed sufficiently long, such that SUMO will not request more values than there are stored in the buffer. This is accomplished by setting the simulation time to a fixed safe value, but might be set dynamically in relation to the buffer size in the future.

The braking control system is a *bang-bang* controller [12] (the brakes are either completely pressed or completely released). Although it is possible to apply different braking forces depending on the position of the braking pedal, a function to determine such position is unavailable in the current TraCI4Matlab documentation. It would also increase the execution time substantially since the Simulink simulation would have to be reinitialized every time the position of the brake pedal changes. On the other hand, SUMO is able to manage this system with good results so this compromise between difficulty, performance and functionality seems like a good trade-off. Alternatively, the simulation of progressively pressing and releasing the brake pedal could be implemented as a part of the braking model if that was desired. In order to detect when the brakes are applied, the brake light signal is employed as the trigger, since it will be activated whenever the brake pedal is pressed.

It is important to note that if the velocity in a vehicle in SUMO is set to 0, SUMO will trigger the emergency brake. In order to avoid this situation, the adapter returns the control to SUMO if the speed is set to a value lower than $0.1mm/s$. This threshold can be changed if desired. After returning the control to SUMO, it stops the car using its original braking system.

2) Settings: In order to run a custom simulation with the adapter, several steps are needed. First of all the SUMO configuration file and the Simulink model should be changed to fit the desired files. Then, if SUMO must be called with some special parameters or using the SUMO GUI, this should be changed in the line that runs SUMO. After that the only other necessary change is adding the necessary initialization values for the Simulink simulation in the designated section.

3) Engineering Decisions: The approach selected in this paper is different to the one shown in [9]. In this implementation, TraCI4Matlab was chosen since it is simpler and more direct than the approach shown in said paper, as seen in Figure 6.

The downside of this approach is that according to [11] the development process of this API is based on reverse engineering of the TraCI for the Python API and it was done quickly and without performance in mind so that might lead to some features not working or underperforming. That said, its application has proven to be easy and adequate for this task.

[3]http://sumo.dlr.de/wiki/Specification#Driving

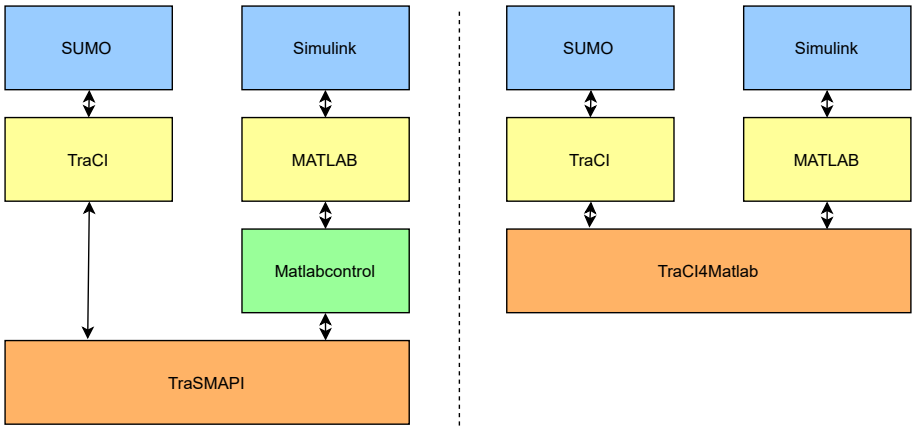


Fig. 6: Comparison of architectures in [9] (left) and this paper (right).

IV. Evaluation

The testing and evaluation has focused on the part that could have made the greatest difference. Since slamming the brakes would result in equivalent deceleration, the test scenario is designed to require a more precise management of the brakes. This scenario, as explained in Section III-A, tries to stop the car at 450 meters from the origin. Some base results from a completely default SUMO simulation (without using an external braking model or the adapter) are presented and compared with the obtained results.

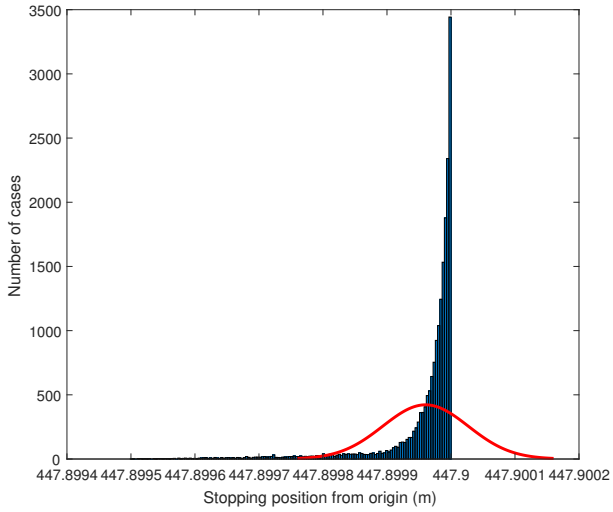
When setting up a vehicle in SUMO there is a parameter called "sigma" which determines how good the driver is. It can be set with any number from 0 to 1, being 0 a perfect driver and 1 a terrible one. Since the simulations with a sigma value of 0 resulted always in the same result, the decision was made to set sigma values of 0.5 and 1 in the simulations.

The results of the original braking system by SUMO are taken as a reference to compare the results of the custom simulations. On the other hand, the results of the simulations with the prototype braking model should show the impact of the adapter in the simulation since, as said in Section III-B, this model simulates the exact mathematical model employed by SUMO to calculate the braking.

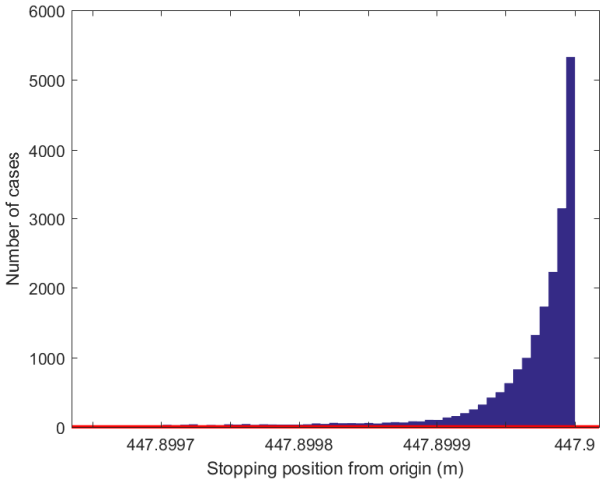
A. *Histograms of the stop position*

All the histograms show the distribution of the braking position of the vehicle in 20000 simulations.

Notably, in Figure 7b, there are 10 points between 428.8202 and 447.4959 meters not shown to increase readability. Both Figures 7a and 7b present sufficiently similar results. As it can be seen, most of the results are registered at the 447.9 meters mark. SUMO will never go over that threshold. This feature was discovered during the earliest tests that were made; when the sigma value of 0 was also tested and all the results returned that exact value. Being this graph representative of an average driver, all the vehicles where able to stop near the said limit. All the other results are tightly grouped around this point (without going over it).

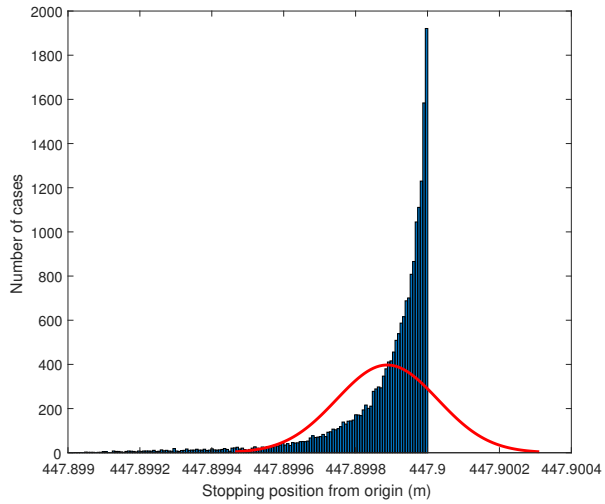


(a) Original

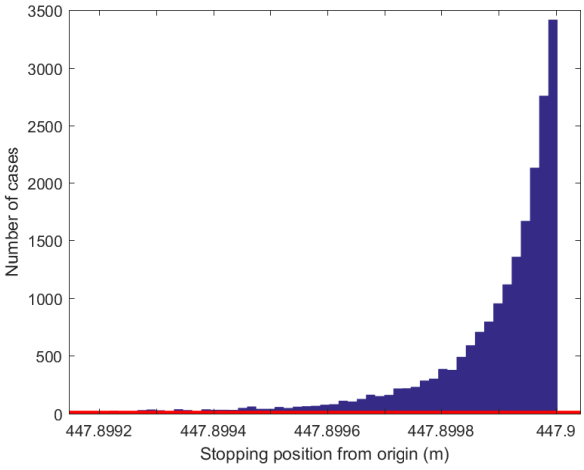


(b) Prototype model

Fig. 7: Comparison of histograms with different braking systems, $\Sigma = 0.5$.



(a) Original



(b) Prototype

Fig. 8: Comparison of histograms with different braking systems and $\Sigma = 1$.

In Figure 8b there are 14 points between 447.4244 and 447.4945 meters not shown to increase readability. The results are similar to the ones in Figure 7. As expected, the results are less tightly grouped due to the sigma value of 1, simulating a very bad driver. Again, there is not one single case where the vehicle stops after the 447.9 meters limit that is set by SUMO. This shows how SUMO is able to manage the use of this braking system without loosing control of the vehicle. The dispersion pattern in both Figures 8a and 8b is seemingly similar (considering the random nature of the experiment) and shows how the behaviour of the simulation with the adapter closely resembles a simulation without it.

V. Conclusion

This paper showed how replacing SUMO's mathematical model for braking by a custom one can be achieved and explains how the adapter is designed. The paper also revealed how the adapter behaves in a different way and the impact that it causes in the simulation. This work provided both a method to employ custom braking models in SUMO as well as overview of its performance. Future work will focus on i) further test scenarios to demonstrate that similarity between old and new brakes are independent of the scenario, ii) applying different parameters in the Simulink model for assessing their individual impact, and iii) an analysis alongside the software and hardware derivatives of the BBW system to determine, if flaws (be they functional or performance related) found in one implementation can be observed in the other.

Acknowledgments

The work in this paper was partly supported by the Knowledge Foundation (KKS) through the project 20130085 (TOCSYC), and by the European Union and VINNOVA under grant 692529-2 (SafeCOP).

References

[1] M. B. Daniel Krajzewicz, Jakob Erdmann and L. Bieker, "Recent development and applications of sumo simulation of urban mobility," in *International Journal On Advances in Systems and Measurements*, vol. 5, no. 3,4, 2012, pp. 128–138.

[2] C. Rhodes and S. Djahel, "Trader:traffic light phases aware driving for reduced traffic congestion in smart cities," in *2017 International Smart Cities Conference (ISC2)*, Wuxi, China, Sept 2017, pp. 1–8.

[3] S. Bhargava, K. Prakasha, and I. Sinha, "Predicting traffic density and increasing fuel efficiency in vehicles using secure vehicular networks," in *2017 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, Jan 2017, pp. 1–3.

[4] N. Müllner, S. Khan, M. H. Rahman, W. Afzal, and M. Saadatmand, "Simulation-based safety testing brake-by-wire," in *2017 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, Tokyo, Japan, March 2017, pp. 61–64.

[5] Y. Mingfei, "Brake-by-wire system design and simulation," in *2012 International Conference on Computer Science and Electronics Engineering*, vol. 3, Hangzhou, China, March 2012, pp. 248–251.

[6] W. M. Griggs and R. N. Shorten, "Embedding real vehicles in sumo for large-scale its scenario emulation," in *2013 International Conference on Connected Vehicles and Expo (ICCVE)*, Dec 2013, pp. 962–963.

[7] E. Martín Campaña, Brake-by-wire in SUMO: A Custom Solution for Braking Simulation, Master of Science Thesis, School of Innovation, Design and Engineering, Mälardalen University, 2018.

[8] D. Chowdhury, L. Santen, and A. Schadschneider, "Statistical physics of vehicular traffic and some related systems," *Physics Reports*, vol. 329, no. 4, pp. 199 – 329, 2000.

[9] J. Macedo, G. Soares, Z. Kokkinogenis, D. Perrotta, and Rosaldo J F Rossett, "A Framework for Electric Bus Powertrain Simulation in Urban Mobility Settings: coupling SUMO with a Matlab/Simulink nanoscopic model," in *1st SUMO User Conference*, BerlinAdlershof, Germany, 2013, pp. 95–102.

[10] J. Macedo, Z. Kokkinogenis, G. Soares, D. Perrotta, and R. J. F. Rossetti, "A hla-based multi-resolution approach to simulating electric vehicles in simulink and sumo," in *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, The Hague, Netherlands, Oct 2013, pp. 2367–2372.

[11] A. F. Acosta, J. E. Espinosa, and J. Espinosa, "Traci4matlab: Enabling the integration of the sumo road traffic simulator and matlabÂő through a software re-engineering process," in *Modeling Mobility with Open Data*, M. Behrisch and M. Weber, Eds. Springer International Publishing, 2015, pp. 155–170.

[12] S. Chen, J. Zhao, and J. Qian, "A design method of bang-bang and pid integrated controller based on rough set," in *Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*, vol. 2, Haikou, China, Aug 2007, pp. 529–533.

Bit Error Rate Mitigation in VANETs Using FFH-OFDM Pre-coding Approach

James Ntaganda
Electrical and Electronics Engineering
College of Science and Technology
University of Rwanda
Kigali, Rwanda

Etienne Ntagwirumugara
Electrical and Electronics Engineering
College of Science and Technology
University of Rwanda
Kigali, Rwanda

Richard Musabe
School of ICT
College of Science and Technology
University of Rwanda
Kigali, Rwanda

***Abstract*—VANET (Vehicular Ad-hoc Network) is based on IEEE 802.11p standard. It is designed for Vehicle to Vehicle (V2V) and Vehicle to Infrastructure (V2I) communication systems purposely for safety, infotainment and traffic management. Nevertheless, due to mobility of communicating nodes, V2V and V2I networks are susceptible to higher errors rates. This is because of Doppler effects on signals caused by relative velocities between communicating vehicles (V2V) or between vehicles and fixed receivers along the pathways (V2I). This paper investigates and proposes the use of fast frequency hopping orthogonal frequency division multiplexing (FFH-OFDM) pre-coding approach to mitigate higher BER associated with Doppler spread in VANETs. FFH-OFDM communication system and Doppler spread channel are modeled and simulated using MATLAB MEX functions and Simulink subject to VANETs standard parameters. BER curves are used as performance metrics for comparison between conventional OFDM and FFH-OFDM models. Under same Doppler spread, simulation results depicted better error performance for FFH-OFDM Pre-coding approach compared to conventional OFDM.**

Keywords—FFH-OFDM, VANETs, Doppler Spread, BER

I. INTRODUCTION

IEEE 802.11p standard for VANETs has come with opportunities and challenges [1]. It is an amended version of IEEE 802.11a standard, and was developed for V2V and V2I communications [2]. Its major applications include road safety, jam management and infotainment, designed to work in the range of 5.85 to 5.925GHz bandwidth and with a bandwidth of 10 MHz per channel [7]. With this bandwidth, it can handle most of broadband communication, including real time video communication such that video can be sent to monitoring stations via V2I links [2, 13].

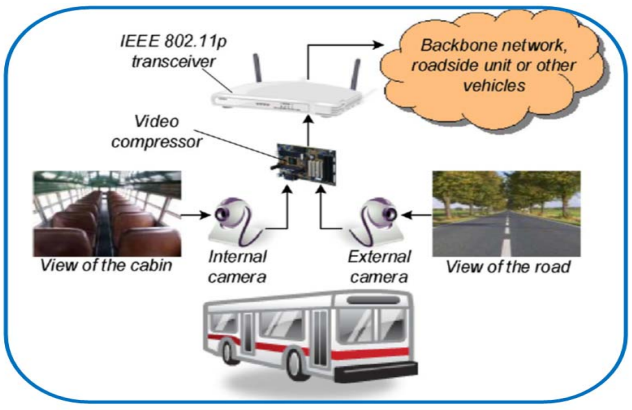


Fig. 1. Video transmission over VANETs [2].

In 802.11p, the bandwidth of 802.11a [12] was halved to make the signal more robust against fading and to increase the tolerance for multipath propagation effects of signals in a vehicular environment [12]. Despite these attempts including other spectral shaping techniques [36], aimed at making VANETs more robust, VANETs remains more susceptible to higher error rates at receiver end. This is due to Doppler spread arising from mobility of vehicles past or towards each other or fixed communication network infrastructure. In this paper, we investigate and propose the use of FFH-OFDM pre-coding approach in VANETs to reduce BER arising from Doppler spread effects. This investigation and proposal is based on; (a) Recent research works [7], which call for further enhancement of VANETs and (b) Existing concepts [5, 6], which established that with FFH-OFDM, a transmitter can alter the modulated frequency according to a predefined "hopping" configuration .Thus, by combining these two schools of thoughts, we aim at mitigating BER affecting VANTEs using FFH-OFDM pre-coding approach. The benefit of FFH-OFDM is that the signal can see a dissimilar channel and a dissimilar set of affected signals for each hop. This evades adversity of deterioration of communication at a particular frequency caused by signal fading or due to specific interferer. The “hopping” configuration is in such a way that the frequency hopping interval is equivalent to an integer number [8], which is also equivalent to time of OFDM symbols. Advantage of this is that even when there is a case of a signal-transmitting sub carrier affected and undergoes intolerable symbol spread and fading at a certain OFDM symbol time, the hopping pattern presents a chance of getting the next symbols which are not affected at the next integer-number time. Hence, by averaging the gain of the frequency diversity, BER can be mitigated at the receiving end. The rest of the sections are organised as follows; section II presents VANETs at glance, previous work related to it as well as challenges and opportunities for research. In Section III, conventional OFDM and FFHOFDM models are presented. In section IV, simulation results are depicted and discussed while section V concludes the paper.

II. VENETS AT GLANCE

VANETs or sometimes called Wireless Access for Vehicular Environments (WAVEs) is a wireless network in which On Board Units (OBU) are deployed in V2V or vehicle to Road Side Units RSU [3, 9, 20, 21]. RSU can as well communicate with centalised servers, monitoring station and trusted authorities such as police stations [9]. Contrary to IEEE 802.11a with 20MHz channel bandwidth, IEEE 802.11p halves the channel bandwidth to be 10MHz. It also operates with in the frequency range of 5.850-5.925 GHz [7, 9, 33,39]. This avails the entire bandwidth capacity of about 75 MHz The spacing in carriers are also halved to (0.15625

MHz) from 802.11a (0.3125 MHz) [10]. On the other hand, the symbol length is doubled. This results into halving transmission rates. In IEEE 802.11p, BPSK, QPSK, 16 QAM and 64 QAM modulation scheme are deployed for efficient packet transmission. Four (4) pilot signals (-21,-7, 7, 21) are slotted in between subcarriers [34]. IEEE 802.11p also specifies an OFDM physical layer that employs 64 subcarriers. 52 out of the 64 subcarriers are used for actual transmission consisting of 48 data subcarriers and 4 pilot subcarriers [10]. Based on these modifications, academia and industry have presented a number of reach works, some of which highlighted challenges [7] as well as opportunities [16, 22, 23, 26, and 32].

A. VANETs Usage, Related Research work and Challenges

Research works [3, 15, 17, 18, 22, 25, 27, 26, 28, 29, 30, 30 and 37] highlighted applicability and opportunities of VANETs. However, Abdeldime M.S. et al [7], presented a list of obvious challenges associated with VANETs and left most of them open for investigation and alleviation. The prominent challenge cited in their presentation [7] was high error rates caused by Doppler spreading. They pointed out that V2V and V2I communication systems are prone to much faster fading, Doppler frequency spread as well as higher multi-path delay spread compared to other wireless systems. The fundamental questions they posed and left open for research were:

(a) "How can we mitigate the Doppler frequency spread in order to grantee low BER and network reliability?"

(b) "Based on the constraints of existing solutions, can we look for unconventional solutions? Or can we look for means and methods to adjust transmitter and receivers according to given Doppler equations parameters?"

(c) "Based on the how channel response varies because of traffic of the vehicles in diverse environments? Can we implement an entire model for VANET channel?"

By summing up [7], the challenge is how to find an appropriate system model that can make vehicular network achieve low BER, high reliability and without much complexity overhead in transreceiver design. Therefore, by tapping into the inherent nature of FFH Scheme, this paper based on existing literature [5, 6], to investigate and finally proposes BER mitigation using FFH-OFDM Pre-coder approach. The benefit of FFH-OFDM is that the signal can see a dissimilar channel and a dissimilar set of spread signals for each hop. This evades adversity of deterioration of communication at a particular frequency caused by signal fading or due to specific interferer.

III. CONVETIONAL OFDM TRANSRECEIVER DESIGN

A number of authors [4, 38] have provided detailed architectures using OFDM. Salem Bindhaiq et al [4] in their paper, made a thorough presentation of OFDM systems and a detailed analysis of Doppler shift effects on OFDM-based communication systems. In spite of OFDM being robust to inter carrier interference (ICI) and intersymbol interference (ISI), in their results [4], a Doppler shift of 100Hz indicated a clear deterioration of BER performance. Adding to concerns in [7], there is a need to mitigate such a challenge. In conventional OFDM scheme, the subcarrier used for transmission of a modulation symbol is fixed for the duration of the OFDM symbol and employs Inverse Fast Fourier Transform (IFFT) during transmission. In such a scheme, OFDM system is designed to carry symbols of data, S (k).

$$S(k) = (s_1, s_2, s_3, \ldots, s_N)^T$$

With this, the number of subcarriers is equivalent to the size of IFFT. The complex vector V (n) is obtained in time domain at the output off IFFT, as in (1).

$$V(n) = (v_1, v_2, v_3.., v_N)^T$$

$$V(n) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} S[k] \mathrm{e}^{j\pi(\frac{2nk}{N})}; n,k = 0,1,\ldots,(N-1) \quad (1)$$

On the other side, upon signal reception, the complex baseband transmitted symbols can be recovered using equation 2.

$$S(k) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} V(n) \mathrm{e}^{-j(2\pi nk/N)}, \; 0 \le n,k \le N-1 \quad (2)$$

For simplicity, we shall represent exponential component in the rest of this paper as (3) [8].

$$[W_{OFDM}]_{n,k} = \mathrm{e}^{j(2\pi nk/N)} \quad (3)$$

Thus, substituting (3) in (1), the transmitted symbols are presented in (4) in a matrix format.

$$V(n) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} [W_{OFDM}]_{n,k} S(k), \quad 0 \le n,k \le N-1 \quad (4)$$

In both equations (1) and (2), the transmitted complex vector '**V (n)"** and recovered symbols of data "**S (k)"** are column vectors. Thus, **(.)^T** denotes transposition operation. From equation (1), $[W_{OFDM}]_{n,k}$ in (3) represents an NxN matrix of conventional IFFT.

A. FFH-OFDM Concepts and Related Work

Contrary to conventional conditional OFDM, FFH-OFDM is not an old concept. It was first coined by T. Scholand et al [5] in their Electronics Letter and used by [19, 35] in their models. Achievement of this letter was the formation of a shuffled matrix of conventional IFFT and later termed FFH/OFDM. The same scuffled matrix was employed by F. Berens, A. et al [6], in their Electronics letter as well and its aim was to design low complex UWB communication systems with ultra high data rates. Until today, no any published work that exploited the shuffled matrix presented in [5, 6, 8], to mitigate the high error rates in VANETs despite the obvious possibilities. In spite of contributions in [5, 6], their work only focused performance of FFH/OFDM subject to ETSI/ITU channel models [39]. Although it was concluded that FFH/OFDM has better performance at high speeds, authors [5, 6] did not tailor it any area of practical application. The effects of Doppler spread were not considered and simulations in [5, 6] based

on 1.9GHz carrier frequency, contrary to VANETs with carrier frequency of 5.9GHz. But, it is worthy acknowledging that their conclusions [5, 6] led to investigative research in this paper. Having found that FFH/OFDM performance is better at higher speeds and given the mobility of communication terminals in VANETs, and based on concerns pointed out in [7], their conclusions called for a thorough research on whether FFH-OFDM would replace conventional OFDM-based IEEE 802.11p. Thus, this paper focuses on a comparative performance analysis of both conventional and non conventional (FFH) OFDM schemes subject to VANETs. That is to say, contrary to previous research, this paper uses 5.9 GHz as the carrier frequency with 10MHz bandwidth per channel and emphasises on Doppler effects. In order to fully realise the deployment of FFH-OFDM in VANETs, pilot symbols were used as well.

B. FFH-OFDM Transmitter Design Model

Use In FFH-OFDM pre-coding approach for VANETs, equation (1) can be modified by an NxN matrix element, presented in (5) [5, 6].

$$e^{\left\{\frac{j2\pi n[\psi]_{n,k}}{N}\right\}} \quad (5)$$

It follows that the transmitted vector symbols are modified as presented in equation (6)

$$V_{ffh-ofdm}(n)=\frac{1}{\sqrt{N}}\sum_{k=0}^{N-1}S(k)e^{j\frac{2\pi n[\psi]_{n,k}}{N}};n,k=0,1,...,(N-1) \quad (6)$$

Based on such modifications, and for simplicity, equation (6) can also be represented in matrix format as in equation (7).

$$V_{ffh-ofdm}(n)=\frac{1}{\sqrt{N}}\sum_{k=0}^{N-1}[W_{FFH-OFDM}]_{n,k}S(k);n,k=0,1,...,(N-1) \quad (7)$$

Case 1: Non cyclic shifting hopping

In this case, entries of a modifying matrix are arranged in an ascending order [8], such that

$$[\psi]_{n,k}=\begin{bmatrix}0 & 1 & \dots & (N-1)\\ 0 & 1 & \dots & (N-1)\\ 0 & 1 & \vdots & (N-1)\\ 0 & 1 & \dots & (N-1)\end{bmatrix},\ n,k=0,1,....(N-1) \quad (8)$$

Consequently, using (8), (5) and (3), a non-hopping IFFT similar to the one employed in conventional OFDM is presented in (9) and the transmitted complex symbols depicted in a matrix format [5, 6, 8] presented in (10)

$$[W_{OFDM}]=\begin{bmatrix}1 & 1 & 1 & \dots & 1\\ 1 & e^{j2\pi/N} & e^{j4\pi/N} & \dots & e^{j2\pi(N-1)/N}\\ \vdots & \vdots & \vdots & \dots & \vdots\\ 1 & e^{j2\pi(N-2)/N} & e^{j4\pi(N-2)/N} & \dots & e^{j2\pi(N-1)(N-2)/N}\\ 1 & e^{j2\pi(N-1)/N} & e^{j4\pi(N-1)/N} & \dots & e^{j2\pi(N-1)(N-2)/N}\end{bmatrix} \quad (9)$$

$$\begin{bmatrix}V(1)\\ V(2)\\ \vdots\\ V(N)\end{bmatrix}=\begin{bmatrix}1 & 1 & 1 & \dots & 1\\ 1 & e^{j2\pi/N} & e^{j4\pi/N} & \dots & e^{j2\pi(N-1)/N}\\ \vdots & \vdots & \vdots & \dots & \vdots\\ 1 & e^{j2\pi(N-2)/N} & e^{j4\pi(N-2)/N} & \dots & e^{j2\pi(N-1)(N-2)/N}\\ 1 & e^{j2\pi(N-1)/N} & e^{j4\pi(N-1)/N} & \dots & e^{j2\pi(N-1)(N-2)/N}\end{bmatrix}*\begin{bmatrix}S(1)\\ S(2)\\ \vdots\\ S(N)\end{bmatrix} \quad (10)$$

Case 2: Cyclically shifted hopping

In such a case, a handy **modulo** operation in (11) permits cyclic hopping with in OFDM symbols employed [5, 6, and 8].

$$[\psi]_{n,k}=\text{mod}[(f_n+k),(N-1)]=\begin{bmatrix}0 & 1 & \dots & (N-1)\\ 1 & 2 & \dots & 0\\ \vdots & \vdots & \vdots & \vdots\\ (N-1) & 0 & \dots & (N-2)\end{bmatrix};n,k=0,1,...(N-1) \quad \text{Equation(11)}$$

In order strike the balance between "hopping" activity and design complexity, the "hopping" is done in such a way that it "hops" in a fixed pattern. This is achieved by introducing a cyclic fixed FFH [8] and in such situation; the unitary matrix; $[U_{FFH\text{-}OFDM}]_{n,k}$ represented in (12) is obtained by introducing a unitary circulant matrix element [5, 6, and 8].

$$[U_{FFH-OFDM}]_{n,k}=([W_{OFDM}]_{n,k})^{H}*\Delta*[W_{OFDM}]_{n,k} \quad (12)$$

The $()^{\mathbf{H}}$ represents a Hermitian operation while **Δ** represents the unitary circulant matrix. To achieve the right circulant hoping patterns, matrix Δ is unitary and diagonal [5, 6, 8].

$$\Delta_{n,k}=e^{(j2\pi n.\frac{n}{N})\delta(n,k)};n,k=0,1,N-1 \quad (13)$$

By substituting the values of **n** and **k** in equation (13), the right circulant matrix **Δ** is represented in (14).

$$\Delta=\begin{bmatrix}1 & 0 & \dots & 0\\ \dots & e^{j2\pi/N} & \dots & 0\\ \dots & \dots & \dots & 0\\ 0 & 0 & 0 & e^{j2\pi(N-1)/N}\end{bmatrix} \quad (14)$$

As a result, IFFT matrix is represented in a shuffled version [5, 6, 8] of conventional IFFT at transmitter end in (16) and the transmitted FFH-OFDM is presented in fig 2.

$$[W_{ffh-ofdm}]=\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ e^{j2\pi/N} & e^{j4\pi/N} & e^{j6\pi/N} & \cdots & 1 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ e^{j2\pi(N-2)(N-2)/N} & e^{j2\pi(N-2)(N-1)/N} & 1 & \cdots & e^{j2\pi(N-2)(N-3)/N} \\ e^{j2\pi(N-1)(N-1)/N} & 1 & e^{j2\pi(N-1)/N} & \cdots & e^{j2\pi(N-1)(N-2)/N} \end{bmatrix} \quad (15)$$

$$V_{ffh-ofdm}(n)=\Delta^{*}[W_{OFDM}]_{n,k}\,S(k) \qquad (16)$$

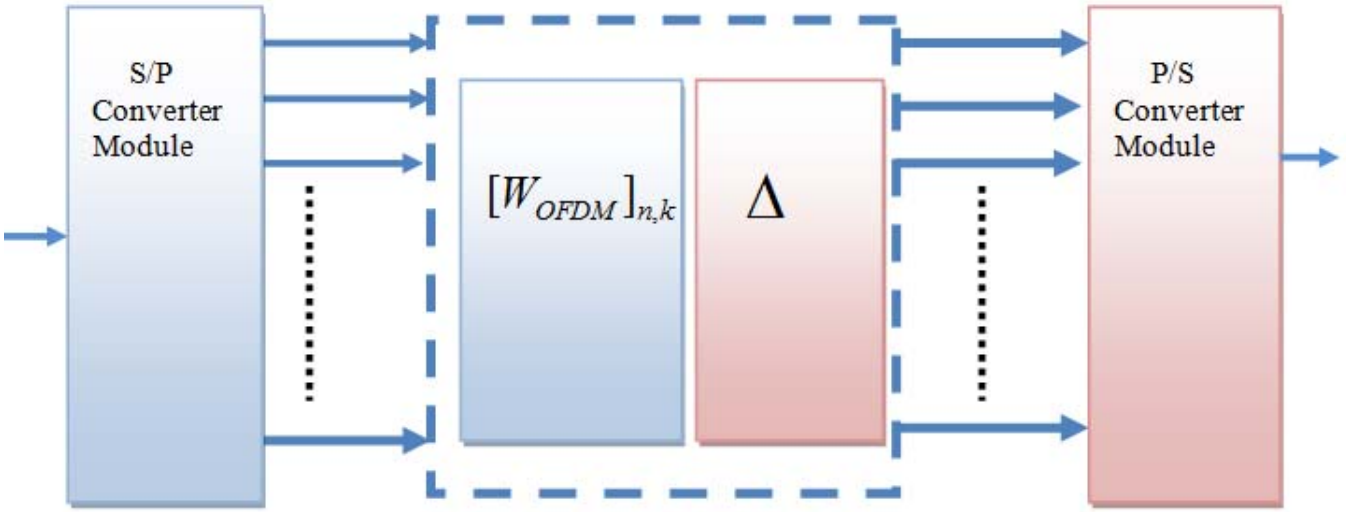


Fig. 2. Representation of FFH unit

C. FFH-OFDM Signal Reception and Equalization

If $V_{ffh\text{-}ofdm}(n)$ is the FFH-OFDM time domain transmitted signal, transmitted via the channel, the transmitted signal in frequency domain is discrete Fourier transform (DFT) of equation (13), obtained by multiplying (16)by Hermitian matrix.

$$\left[[W_{OFDM}]_{n,k}\right]^{H}$$

This multiplication leads to a unitary matrix

$$[U_{FFH-OFDM}]_{n,k}$$

The same matrix is depicted in (18)[5,6,8].Assuming that the noise density, **N0** is a priori known to the receiver, the additive white Gaussian noise (AWGN) with double-sided has a spectral noise density of **N0/2**[5,6]. Thus, the noise vector in frequency domain that impairs the signal is represented in equation (19). Considering an NxN diagonal channel matrix, H_c, in which each entry is a mobile radio transfer function measured at the center frequency of each subcarrier, the received vector, R_V is represented in frequency domain (20).

$$[U_{FFH-OFDM}]_{n,k}=([W_{OFDM}]_{n,k})^{H}*\Delta*[W_{OFDM}]_{n,k} \quad (18)$$

$$n_c=(n_1,n_2,n_3,\ldots,n_N)^{T} \qquad (19)$$

$$R_v=[H_C*[U_{FFH-OFDM}]_{n,k}*s_k+n_c$$

$$R_v=[H_C*\left[[W_{OFDM}]_{n,k}\right]^{H}*\Delta*[W_{OFDM}]_{n,k}]*s_k+n_c \quad (20)$$

If we apply an inverse of a diagonal matrix, Δ to vector, $[V_{ffh\text{-}ofdm}]_n$, the orthogonality of FFH-OFDM symbols can be recovered. Furthermore, because in fast Frequency hopping the Hermitian of IFFT matrix is unitary in nature, then we get $\Delta^{-1}=\Delta^{*}$ [5, 6, 14]. For received signal equalisation, the most commonly used channel equalizers in OFDM-based communication systems are MMSE and ZF. Both equalizers require NxN block linear and diagonal equalizer matrices. For there is no need to reinvent the wheel, and based on the fact that the specific aim of this paper is to compare the performance of FFH-OFDM vs Conational OFDM subject to VANETs, this paper used ZF equalizer because of its simplicity in design and simulations[24,39]. Thus, the estimated equalizer output is represented in equation (21).

$$\hat{S}_{ZF}=\left[[W_{OFDM}]_{n,k}\right]^{H}.\Delta^{*}.\left[[W_{OFDM}]_{n,k}\right].M_C^{-1}.R_v \qquad (21)$$

D. Doppler effects and channel models

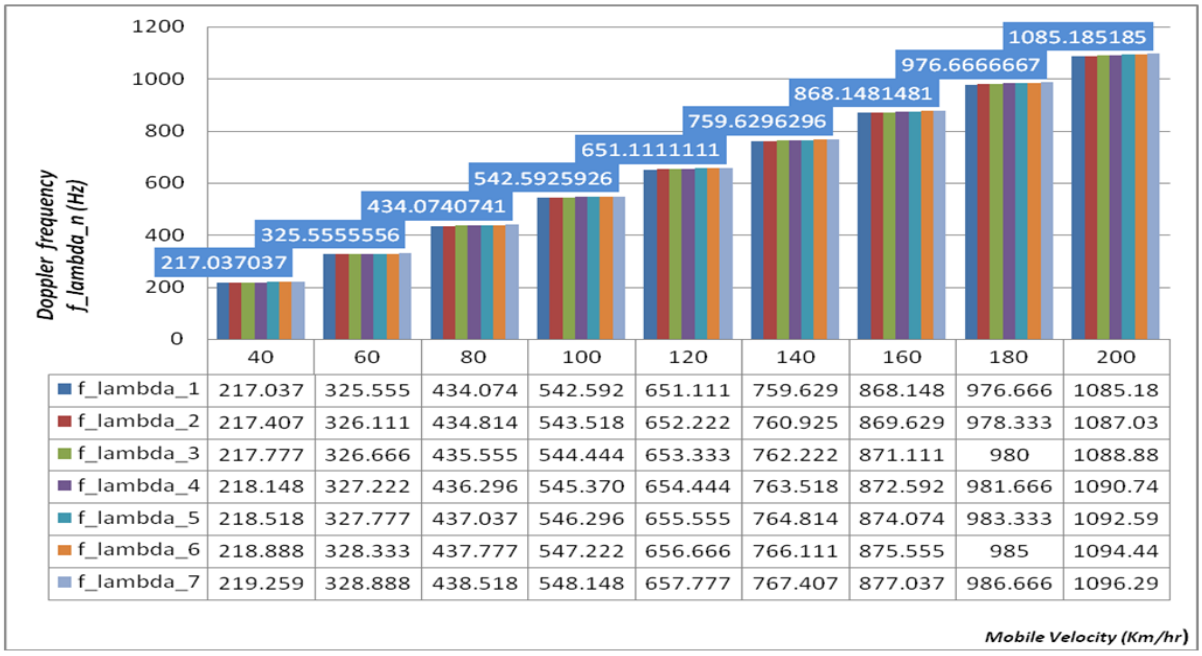
OFDM systems offer a number of advantages. One of them is that it is able to change frequency selective channel to a number of flat fading channels. This is archived by picking suitable number of sub-carriers for OFDM signals. In this situation, sub-bandwidths of the OFDM signals should be smaller than the coherence channel bandwidth and in order to reduce the design complexity, a single tap equalizer is deployed. However, MC-OFDM is prone to frequency offsets a raising from Doppler spread. That is to say, if we scale down the whole bandwidth into many smaller sub-bandwidths, they become comparable with Doppler frequency arising from time varying channel, especially in VANETs. The Doppler frequency can be represented in terms of mobile velocity (v) and carrier wavelength (λc). The relationship is given by (22)

$$f_\lambda = \frac{v}{\lambda_c}\cos(\theta) \quad (22)$$

,

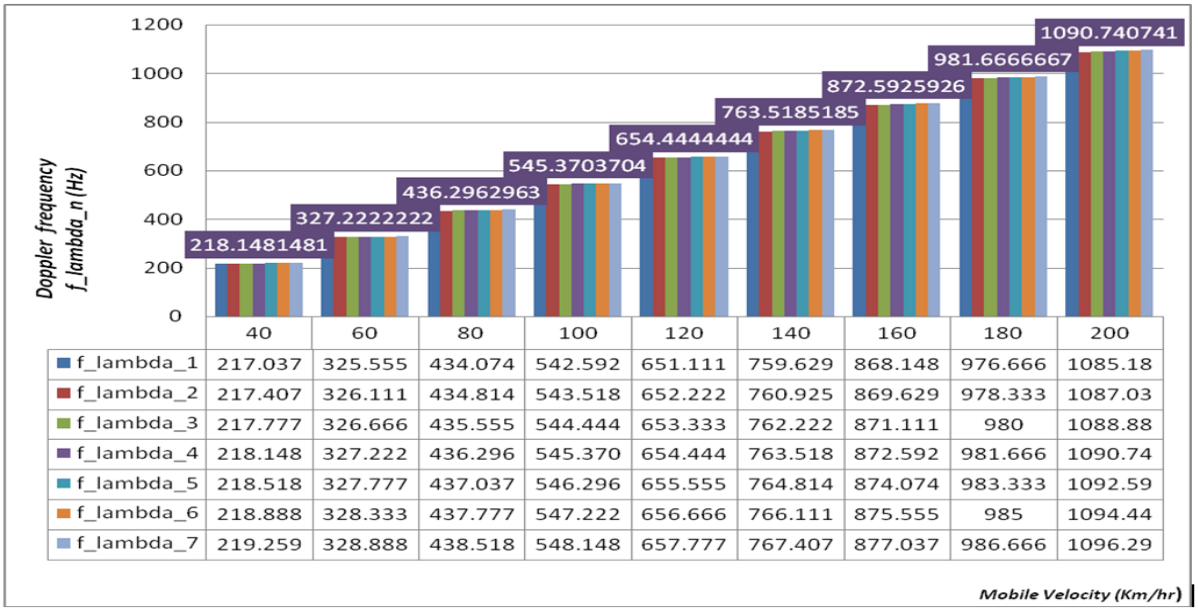
In this equation, θ is the incident angle. Doppler spread leads to dissimilarity in the CIR during a single OFDM symbol broadcast. This also leads to non-zero correlation of carriers due to loss of orthogonality of the OFDM subcarriers, hence ICI and increased BER. In this paper, Doppler spread was modeled by Jakes Channel Model (JCM) by Zeng and Xiao [40, 41].

Using the normalised equation version of (22), for simulation purposes, a range of Doppler frequencies are presented in **fig. 3, 4 and 5.** In the same figures, **f_lambda_1** up to **f_lambda_7** represent Doppler frequencies corresponding to the carrier frequencies of respective VANET channels **172** to channel **184** respectively.



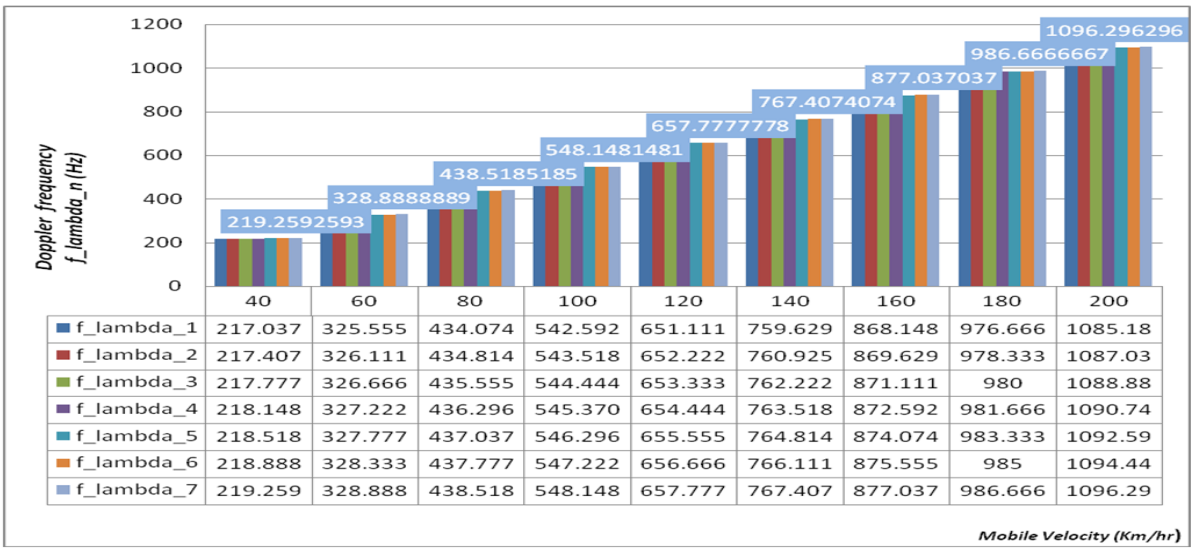
	40	60	80	100	120	140	160	180	200
f_lambda_1	217.037	325.555	434.074	542.592	651.111	759.629	868.148	976.666	1085.18
f_lambda_2	217.407	326.111	434.814	543.518	652.222	760.925	869.629	978.333	1087.03
f_lambda_3	217.777	326.666	435.555	544.444	653.333	762.222	871.111	980	1088.88
f_lambda_4	218.148	327.222	436.296	545.370	654.444	763.518	872.592	981.666	1090.74
f_lambda_5	218.518	327.777	437.037	546.296	655.555	764.814	874.074	983.333	1092.59
f_lambda_6	218.888	328.333	437.777	547.222	656.666	766.111	875.555	985	1094.44
f_lambda_7	219.259	328.888	438.518	548.148	657.777	767.407	877.037	986.666	1096.29

Fig. 3. Doppler frequency, **f_lambda_1 (217.037-10885.18 Hz)** plotted against Mobile velocity (40-200 Km/hr) respectively, for channel **172** in VANETS



	40	60	80	100	120	140	160	180	200
f_lambda_1	217.037	325.555	434.074	542.592	651.111	759.629	868.148	976.666	1085.18
f_lambda_2	217.407	326.111	434.814	543.518	652.222	760.925	869.629	978.333	1087.03
f_lambda_3	217.777	326.666	435.555	544.444	653.333	762.222	871.111	980	1088.88
f_lambda_4	218.148	327.222	436.296	545.370	654.444	763.518	872.592	981.666	1090.74
f_lambda_5	218.518	327.777	437.037	546.296	655.555	764.814	874.074	983.333	1092.59
f_lambda_6	218.888	328.333	437.777	547.222	656.666	766.111	875.555	985	1094.44
f_lambda_7	219.259	328.888	438.518	548.148	657.777	767.407	877.037	986.666	1096.29

Fig. 4. Doppler frequency, f_lambda_4 (218.148 to 1090.74 Hz) plotted against Mobile velocity (40 to 200 Km/hr) respectively, for channel **178** in VANETS



	40	60	80	100	120	140	160	180	200
f_lambda_1	217.037	325.555	434.074	542.592	651.111	759.629	868.148	976.666	1085.18
f_lambda_2	217.407	326.111	434.814	543.518	652.222	760.925	869.629	978.333	1087.03
f_lambda_3	217.777	326.666	435.555	544.444	653.333	762.222	871.111	980	1088.88
f_lambda_4	218.148	327.222	436.296	545.370	654.444	763.518	872.592	981.666	1090.74
f_lambda_5	218.518	327.777	437.037	546.296	655.555	764.814	874.074	983.333	1092.59
f_lambda_6	218.888	328.333	437.777	547.222	656.666	766.111	875.555	985	1094.44
f_lambda_7	219.259	328.888	438.518	548.148	657.777	767.407	877.037	986.666	1096.29

Fig.5. Doppler frequency, f_lambda_7 (219.259 to 1096.29 Hz) plotted against Mobile velocity (**40 to 200 Km/hr**) respectively, for channel **184** in VANETS

IV SYSTEM SIMULATION RESULTS

During simulations, carrier frequencies of channels 172 to 184 in VANETs were used and a rage of Vehicular speed is considered from 40Km/hr and increased up to 200km/hr. In Fig 3, 4 and 5, respective Doppler frequencies are depicted. As the focus of this paper, conventional OFDM communication systems are compared to FFH-OFDM systems. Comparison is based on numerical simulations and takes into account Doppler spread as the prime cause of higher BER is VANETs. Simulations used flat Rayleigh fading channel and all subcarriers tuned to an equal channel gain. The performance metric for comparison is BER curves. QPSK is employed as a form of modulation. Fig 3, 4 and 5 depict data sets of Doppler frequency associated the carrier frequency of 5.86GHz, 5.87 GHz and 5.92GHz with carrier wave length of 0.051194539m, 0.050933786m and 0.050675676m respectively. The ranges of Doppler frequencies are obtained as velocity varies from 40Km/hr to 200Km/hr. It can be shown in fig 3, 4 and 5 that as speed increases, the Doppler spread also increases, which is why high BER is of great concern in high mobility networks, and needs mitigation approaches. Results in Fig 6 and 7 indicate a better performance for FFH-OFDM compared to conventional OFDM. Another interesting scenario from our simulations is that, Results Fig.7, indicate that at higher speeds (200Km/hr) BER tends to a constant rate even when SNR is increased.

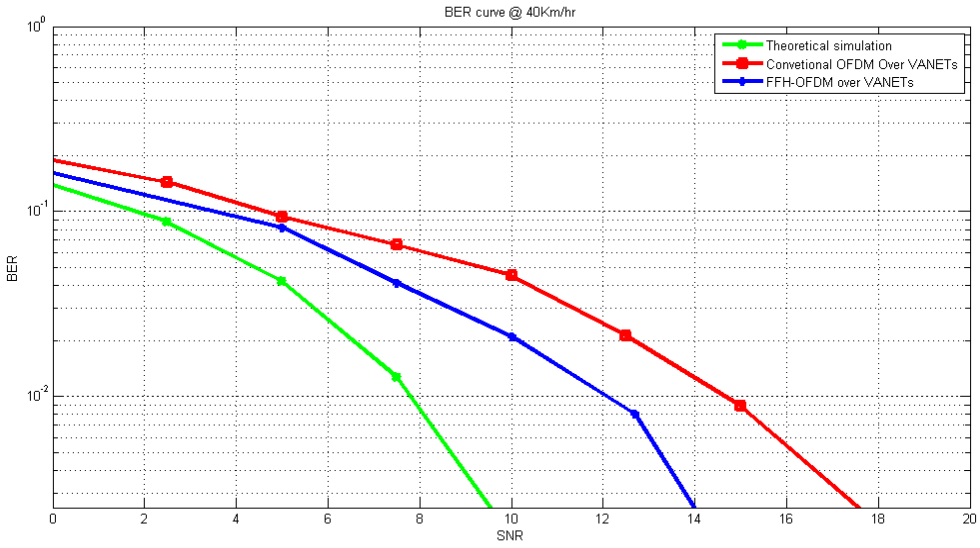


Fig .6. BER Simulation results using the worst case scenario (channel 184, lambda_7) at a 40Km/hr speed

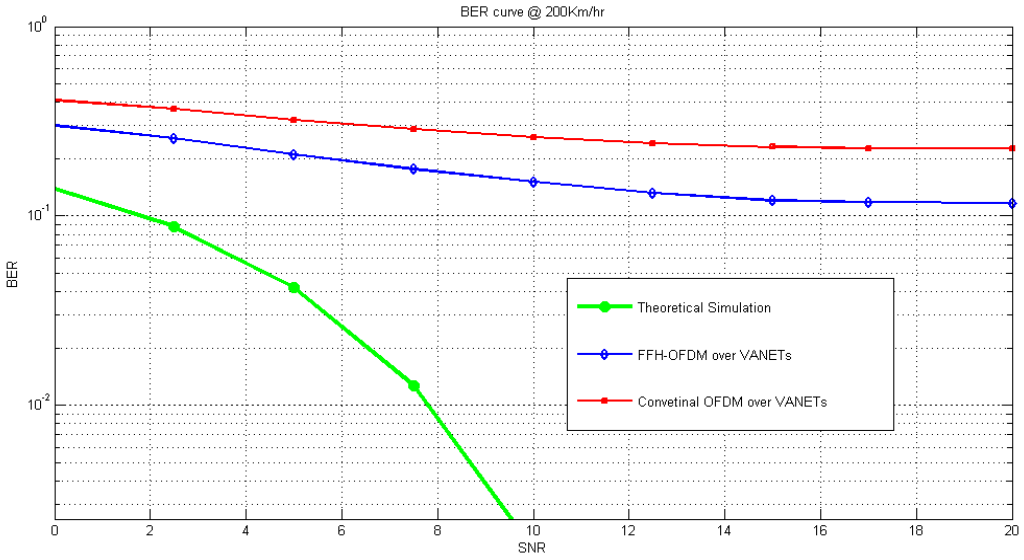


Fig.7. BER Simulation results using the worst case scenario (channel 184, lambda_7) at a 200 Km/hr speed

V. CONCLUSION

This paper investigated the possibility using of FFH-OFDM linear pre-coder to mitigate higher BER in VANETs, caused by Doppler spread, arising from relative velocities in vehicular environment. We present results for 40Km/hr (congested city speed) and 200 Km/hr (extreme highway speed). BER Simulations used a carrier frequency of 5.87 GHz, with carrier wave length of 0.050933786m subject to VANETs. This frequency was chosen as the best candidate to represent all simulations because it lies in the middle, between the seven channels in VANETs. BER is used as evaluation and performance metric. Based on simulation results, FFH-OFDM indicated better performance compared to conventional OFDM, as speed increased beyond 40Km/hr. However, simulations indicated less difference in BER at speeds lower than 40Km/hr. additionally, as speed increased to 200Km/hr and beyond, the probability of error rates did not reduce in both cases, even when SNR was increased. Based on simulation results, it can be concluded that in the ranges of 40Km/hr up to 200Km/hr, FFH-OFDM pre-coder can mitigate errors caused by Doppler spread. However, the remaining question and probably an area that requires further investigation from Hardware designers, is whether the design overhead due to additional blocks (pre-coding block unit) is worthy pursuing, and whether better BER performance for FFH-OFDM can be traded-off to any hardware design complexity caused by additional functional blocks.

REFERENCES

[1] Y.Yang, H. Li, Q.Huang. "Mobility Management in VANET", *22nd Wireless and Optical Communication Conference (WOCC),* 16-18 May 2013, Chongqing, China, doi: 10.1109/WOCC.2013.6676326

[2] E. Belyaev, A.Vinel, A.Surak, M.Gabbouj ,M.Jonsson, K. Egiazarian, "Robust Vehicle-to-Infrastructure Video Transmission for Road Surveillance Applications", IEEE *transactions on vehicular technology,* vol. 64, no. 7, July 2015

[3] T. Sukuvaara, R.Ylitalo, and M. Katz. "IEEE 802.11p Based Vehicular Networking Operational Pilot Field Measurement". IEEE *journal on selected areas in communications/supplement*, VOL. 31, NO. 9, September 2013

[4] S.Bindhaiq, S.Kamilah Syed-Yusof and H. Hosseini, "Performance analysis of Doppler shift effects on OFDM-based and MC-CDMA-based cognitive radios", *International journal of communication systems , 2013, DOI: 10.1002/dac.2497*

[5] T. Scholand, T. Faber, A. Seebens, J. Lee, J. Cho, Y. Cho, H.W. Lee and P. Jung, "Fast frequency hopping OFDM concept", Electronics Letters (Volume: 41, Issue: 13, 23 June 2005) , **DOI:** 10.1049/el:20050644

[6] F. Berens, A. Ru¨egg, T. Scholand and A. Hessamian-Alinejad, " Fast frequency hopping diversity scheme for OFDM-based UWB systems". IET, *Electronics Letters (Volume: 43, Issue: 1, Jan. 4 2007)* 2007 **DOI:** 10.1049/el:20073776

[7] Abdeldime M.S. Abdelgader, W. Lenan, "The physical layer of IEEE 802.11p WAVE communication standard: The specifications and Challenges. *Proceedings of the World Congress on Engineering and Computer Science* 2014 Vol II WCECS 2014, 22-24 October, 2014, San Francisco, USA

[8] Farooq Khan, "LTE for 4G Mobile Broadband: Air Interface Technologies and Performance", *Cambridge University Press*, 2009 pp 55-62

[9] W. Chim, S. M. Yiu, Lucas C. K. Hui and Victor O. K. Li, "Privacy-Preserving Information Gathering Using VANET",2012 http://dx.doi.org/10.5772/53537

[10] G. C. Kiokes, G.Economakos, A.Amditis and N.K. Uzunoglu, " Recursive Systematic Convolutional Code Simulation for Ofdm - 802.11p System and FPGA Implementation Using an ESL Methodology". *12th Euromicro Conference on Digital System Design / Architectures, Methods and Tools*, 2009

[11] Ikbal Chammakhi Msadaa, Pasquale Cataldi and Fethi Filali, "A Comparative Study between 802.11p and Mobile WiMAX-based V2I Communication Networks", *Next Generation Mobile Applications, Services and Technologies (NGMAST), 2010 Fourth International Conference* ,2010

[12] Ó.Puñal, C. Pereira, A.Aguiar and J. Gross, "Experimental Characterization and Modeling of RF Jamming Attacks on VANETs", *IEEE transactions on vehicular technology*, vol. 64, no. 2, February 2015

[13] E.Yaacoub, F.Filali, and A.Abu-Dayya, "QoE Enhancement of SVC Video Streaming Over Vehicular Networks Using Cooperative LTE/802.11p Communications", *IEEE journal of selected topics in signal processing*, VOL. 9, No. 1, February 2015

[14] Zheng Li, Fan Bai, Joseph A. Fernandez, B. V. K. Vijaya Kumar "Tentpoles Scheme: A Data-Aided Channel Estimation Mechanism

for Achieving Reliable Vehicle-to-Vehicle Communications", *IEEE transactions on wireless communications*, vol. 14, no. 5, may 2015

[15] Kunjie Xu, D. Tipper, Yi Qian and P.Krishnamurthy, "Time Dependent Performance Analysis of IEEE 802.11p Vehicular Networks", *2016 in IEEE Transactions on Vehicular Technology, 2016,* ***DOI:*** *10.1109/TVT.2015.2470113*

[16] N.Bouchemal , R.Naja , M. Moubarak, and S.Tohme. "*Modeling and Performance Evaluation of V2I Communications" ,* ***IEEE Conference Publications****, Beirut, Lebanon* 2015, 18 - 25,

DOI: 10.1109/DICTAP.2015.7113164

[17] L.Bernadó, T. Zemen, F.Tufvesson, A F. Molisch, and C. F. Mecklenbräuker, "Time- and Frequency-Varying K-Factor of Non-Stationary Vehicular Channels for Safety-Relevant Scenarios", *IEEE transactions on intelligent transportation systems*, vol. 16, no. 2, April 2015

[18] W.Y.Lin, M.W. Li, K.C. Lan, and C.H.Hsu, "A Comparison of 802.11a and 802.11p for V-to-I Communication: A Measurement Study", *Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, 2011

[19] Y.He, Y. Cheng, Y.Yang, G.Wu, B. Dong, and S.Li, "A Subset-Based Coherent FFH System", *IEEE communications letters*, vol. 19, no. 2, February 2015

[20] M.T.Moreno, J.Mittag, P.Santi, and H.Hartenstein, "Vehicle-to-Vehicle Communication: Fair Transmit Power Control for Safety-Critical Information", *IEEE transactions on vehicular technology*, VOL. 58, NO. 7, September, 2009

[21] W.Xiang, D. Shan, J.Yuan, and S. Addepalli, "A Full Functional Wireless Access for Vehicular Environments (WAVE) Prototype upon the IEEE802.11p Standard for Vehicular Communications and Networks", *IEEE Consumer communication and Network* ,2012

[22] S.Biddlestone, K.Redmill, R.Miucic, and Ü. Özgüner, "An Integrated 802.11p WAVE DSRC and Vehicle Traffic Simulator With Experimentally Validated Urban (LOS and NLOS) Propagation Models", *IEEE transactions on intelligent transportation systems, vol. 13, no. 4, December* , 2012

[23] W.J.Lin, I.C. Liu, and J.S. Lin, "Novel Frequency Offset Estimation for IEEE 802.11p Standard in V2V Channels", *IEEE International Conference on ITS Telecommunications*, 2012

[24] Z. Zhao ,X.Cheng, M.Wen and B. Jiao, Cheng-Xiang Wang, "Channel Estimation Schemes for IEEE 802.11p Standard", *IEEE Intelligent transportation systems magazine*, 2013

[25] H.Noori, B.Badihi and Olyaei, "A Novel Study on Beaconing for VANET-basedVehicle to Vehicle Communication:Probability of Beacon Delivery in Realistic Large-Scale Urban Area using 802.11p", IEEE, 2013

[26] S. E. Carpenter, "Obstacle Shadowing Influences in VANET Safety", *22nd International Conference on Network Protocols*. IEEE ,2014

[27] S. Moser, L. Behrendt , F. Slomka, "MIMO-Enabling PHY Layer Enhancement for Vehicular Ad-Hoc Networks", Wireless Communication and Network. IEEE, 2015

[28] J. Santa, P. J. Fernández, F.Pereñíguez, F. Bernal, A. F. Skarmeta, "A Vehicular Network Mobility Framework: Architecture, Deployment and Evaluation", *IEEE INFOCOM International Workshop on Mobility Management in the Network of future Word*, 2015

[29] A. Skårbratt, and R.Rehammar, "802.11p measurements in reverberation chamber", IEEE 2015

[30] G. Kiokes , G.Economakos , A. Amditis and N. K. Uzunoglu, "A Comparative Study of IEEE 802.11p Physical Layer Coding Schemes and FPGA Implementation for Inter Vehicle Communications", *Modern Traffic and Transportation Engineering Research,* April 2013

[31] K. Bilstrup, E. Uhlemann and E G. Ström, "Urban Bilstrup, Evaluation of the IEEE 802.11p MAC method for Vehicle-to-Vehicle Communication", IEEE, 2008

[32] I.C. Msadaa, P. Cataldi , F. Filali, "A Comparative Study between 802.11p and Mobile WiMAX-based V2I Communication Networks", IEEE 2010

[33] C. Campolo, A.Vinel, A. Molinaro, and Y. Koucheryavy, " Modeling Broadcasting in IEEE 802.11p/WAVE Vehicular Networks", *IEEE COMMUNICATIONS LETTERS*, VOL. 15, NO. 2, February 2011

[34] A. sassi, F.charfi, L. kamoun,Y. elhillali and A. rivenq, "OFDM Transmission Performance Evaluation in V2X Communication", *IJCSI International Journal of Computer Science Issues*, Vol. 9, Issue 2, No 3, March 2012

[35] L. Mailaender "Anti-Jam Communications using Frequency-Hopped OFDM and LDPC with Erasures, *conference paper* " , November 2013

[36] T. H. Pham, I.V McLoughlin, S. A. Fahmy. Shaping Spectral Leakage for IEEE 802.11p Vehicular Communications. *IEEE* ***Conference Publications****, 2014,* p1 - 5, DOI: 10.1109/VTCSpring.2014.7023089,

[37] Nikolajs Agafonovs, Girts Strazdins, Modris Greitans.Accessible, Customizable, High-Performance IEEE 802.11p Vehicular Communication Solution, ***IEEE Conference Publications,2012,*** *127 - 132, DOI: 10.1109/MedHocNet.2012.6257112*

[38] Y. Han; P. Harliman; S. W. Kim. J. K. Kim; C. Kim, "A Novel Architecture for Block Interleaving Algorithm in MB-OFDM Using Mixed Radix System", *EEE Transactions on Very Large Scale Integration (VLSI) Systems, IEEE Journals & Magazines*, 2010. Pg.1020 - 1024, DOI: 10.1109/TVLSI.2009.2018091

[39] Molisch, A., Foerster, J., and Pendergrass, M, "Channel models for ultra wideband personal area networks", *IEEE Wireless Communication.*, 2003, 10, (6), pp. 14–21

[40] Y.R. Zeng and C. Xiao, "Improved Models for the Generation of Multiple Uncorrelated Rayleigh Fading Waveforms", *IEEE Comm.Letters*, vol. 6, no. 6, June 2002.

[41] G. L. Stüber, "Principles of Mobile Communication", 2nd Edition, KluwerAcademic Publishers, New York, Boston, Dordrecht, London, Moscow, 2002.

Framework for Visual Simultaneous Localization and Mapping in a Noisy Static Environment

Jabulani K. Makhubela, Tranos Zuva, and Olusanya Yinka Agunbiade
Faculty of Applied and Computer Sciences
Dept of Information and Communication, Vaal University of Technology
Vanderbijlpark, South Africa
kingsley.makhubela@gmail.com

***Abstract*—Simultaneous Localization and Mapping (SLAM) has seen a tremendous interest amongst research community in recent years due to its ability to make the robot truly independent in navigation. The capability of an autonomous robot to locate itself within the environment and construct a map at the same time, it is known as Simultaneous Localization and Mapping (SLAM). Visual Simultaneous Localization and Mapping (VSLAM) is when autonomous robot employs a vision sensor such a camera to explore the environment. Various researchers have embarked on the study of Visual Simultaneous Localization and Mapping (VSLAM) with excellent results, however the challenge of environmental noise such as light intensity still persist. In this study we propose a framework for Visual Simultaneous Localization and Mapping (VSLAM) that will address the challenge of light intensity in an environment in order to improve the performance of Visual Simultaneous Localization and Mapping (VSLAM) system. In executing of Visual Simultaneous Localization and Mapping (VSLAM) system, we have introduced a filtering algorithm to reduce or limit the effects of noise on images taken from the environment. The outcome of our study is a framework that will enable an autonomous robot to successfully navigate, localize itself and map the environment.**

Keywords—navigation, sensors, vision, illumination variance, Simultaneous Localization and Mapping (SLAM)

I. INTRODUCTION

Deciding on the stance of a moving robot (self-localization) in the surrounding location is one of the great issues of mobile autonomous robots [1]. A mobile robot must know where it is inside a location to explore self-sufficiently and cleverly [1]. Self-localization and knowing the area of different things requires the presence of a model on the environment, and this essential necessity has prompted the development of the Simultaneous Localization and Mapping (SLAM) algorithm in the course of recent decades, where the robot assembles a model piece-wise as it investigates location [2]. This prompts the following "the chicken or the egg" issue: for self-localization, the robot requires a map, yet to construct such a guide, the posture of the robot must be known. An answer for this issue is to assemble the model while registering a posture evaluate, which is known as Simultaneous Localization and Mapping (SLAM) [1].

Simultaneous Localization and Mapping (SLAM) is portrayed as the capacity of an independent robot to create a model of an obscure surrounding and estimate its state in the same time [3]. Simultaneous Localization and mapping (SLAM) research area has been a typical challenge for autonomous robots to explore and achieve mobile operational exercises [4]. As per [5] the Simultaneous Localization and Mapping (SLAM) research area has been invaluable for resolving autonomous robot weaknesses for the research community; it has additionally pulled in numerous research studies and is the fundamental explanation behind robotics accomplishment [6]. The key purpose behind the Simultaneous Localization and Mapping (SLAM) research area's prosperity is that it is superior to others in evacuating the artificial foundations of the area [5].

For an autonomous robot to explore the environment, the larger part of frameworks utilizes an extensive variety of sensors, for example, laser scanner, sonar, acoustic, and so forth. [7], this sensor which are utilized into Simultaneous Localization and Mapping (SLAM) are either named a laser, sonar and additionally vision sensor. Visual Simultaneous Localization and Mapping (VSLAM) it is a capacity of an autonomous robot to explore, model and estimate its location by using administrations of a vision sensor as its fundamental sensor [8]. Visual Simultaneous Localization and mapping (VSLAM) is when autonomous robot inserted with a Vison sensor, for example, monocular, stereo vision, omnidirectional or/and Red Green Blue Depth (RGBD) camera [9].

Park and Song embrace the activity of Vision Sensor into the Simultaneous Localization and Mapping (SLAM) system that can convey more important data on the area than any other sensors [10]. Various Sensors, for example, laser scanner, sonar, acoustic, and so forth that can be substantial and costly [7]. [11] express that the advantages of a Vision Sensor are that they are reduced, exact, low-evaluated, non-obtrusive and inescapable. Cameras as vision sensors are favored, in light of the fact that individuals and creatures appear to explore viably in complicated areas utilizing vision as prime sensor [7].

Driving methodologies and algorithms in Visual Simultaneous Localization and Mapping (SLAM) research area that uses Extended Kalmant Filter (EKF), FASTSLAM, GraphSLAM, and Rao-Blackwellized Particle Filter (RBPF), coSLAM, monoSLAM etc., have been displayed by different researchers [12)]. Various researchers have focused on making the vision Visual Simultaneous Localization and Mapping (VSLAM) procedure to work on all the more difficult conditions [8], However the issue of Environmental noise, for example, light intensity in a static environment has been a one of the issues that causes limits the performance on the Visual Simultaneous Localization and Mapping (VSLAM) [8]. In an attempt to solve this challenge in our study, we will introduce noise filtering algorithm into the Visual Simultaneous Localization and Mapping (SLAM) framework to increase the effectiveness of the system. Experimental performance and simulation will be done using Matlab software and analyses will be carried out using both

qualitative and quantitative methods and experimental outcomes obtained will be compared with existing outcomes from the literature to reveal the system with the best performance.

The outcome of this study will improve Visual Simultaneous Localization and Mapping (VSLAM) and it will enable an autonomous robot application to navigate the static environment where environmental noise exists without causing an error. Visual Simultaneous Localization and Mapping (VSLAM) technique implementation will contribute to autonomous navigation research and reduce human catastrophe while also assisting in production.

II. Literature Review

A. Introduction

For an autonomous robot to localize its position and successfully build a map in unfamiliar location, it needs to follow a method called Simultaneous Localization and Mapping (SLAM) [13]. Visual Simultaneous Localization and Mapping (VSLAM) is an autonomous robot utilizing a camera to navigate the environment [8]. Simultaneous localization and mapping (SLAM) is a common research area of focus for the scientific community. Various research in this field has been carried out with remarkable results [14], however challenges still exists in Simultaneous localization and mapping. This section will review some of the work proposed by these researchers and outline their limitations they face when implementing their method.

A proposed technique by [15] on Stereo Vision Simultaneous Localization and Mapping (VSLAM) for independent mobile robot route in an indoor location. The goal was to outline a framework in which an autonomous robot would solely use a vision sensor for obtaining information and exploring the environment. Their navigation framework contained route and self-localization. The general route progressive system contained of localization, Perform the Region of Interest (ROI), Region of Interest (ROI) Sub Screening, framework mapping ideal way pursuit and way arranging. The standard exercises into their Visual Simultaneous Localization and Mapping (SLAM) route framework was to accomplish 3D depth estimation of the area, Scene examination, ideal way look, constant way arranging and engine speed control. The experimental outcome demonstrated their proposed system was fit for exploring in a pre-arranged paths condition emphatically. However, the data association on the mobile robot turned into a difficult because of varieties in light, specular appearance in environment and conflicting point clouds which happened as a result of the variety in review edges of the camera.

In reference to a research done by [9] for Robust Visual based Simultaneous Localization and Mapping (VSLAM) utilizing planar point's features with the aid of a Red Green Blue Depth (RGBD) sensor. The goal of the study was to propose a technique that would enhance the precision in capturing and recording point's features into the Visual Simultaneous Localization and Mapping (VSLAM). Their proposed method utilized a planar feature to adjust key frames, the proposed technique would begin by expelling planes points from cloud acquired from the Red Green Blue Depth (RGBD) sensor, at that point endeavors enhance the quality of the mined planes, at that point depth estimations of planes would be altered by the parameter of plane models and Random Sample Consensus (RANSAC) structure is then utilized to detect and compare point feature on the individual match of planes. As indicated by [16] Random Sample Consensus (RANSAC) is a general technique for redressing model in the presence of outliers. Their proposed technique additionally contained a graphic end part which executed picture process ego-motion calculation and pose optimization part was for executing global pose graph optimization. The experimental results of their proposed technique on open datasets demonstrated following direction is genuinely exact, high exactness in movement of the two frames with less failure rate. However physical examination outcomes demonstrated dreary structure on floor, light from the roof disturbed the matching process and daylight from window upsets the enlightenment conditions.

In the examination directed by [17], for Visual Simultaneous Localization and Mapping (VSLAM) based Indoor/Outdoor location obstacle evasion. The reason for the investigation was to make an on-broad stereo-vision based mapping framework that would be utilized for path planning and local obstacle evasion in a search and rescue mission. Their framework configuration comprised of three layers which was recognition, mapping also planning and control. In outside set-up the tight edge and wide-edge Vision Simultaneous Localization and Mapping (VSLAM) framework delineated an above exactness of 0.22m mean fault, having an enhanced deviation of under 0.08% and 0.06% of the full course implying that the Vision Simultaneous Localization and Mapping (VSLAM) framework was proficient to deal with mistakes, floating combination computations and able to correct the drifts also. A Vision Simultaneous Localization and Mapping (VSLAM) framework likewise demonstrated positive advance in shutting the substantial circle in an exploratory set-up. The Vision Simultaneous Localization and Mapping (VSLAM) framework showed that it could manage with changing light blended Indoor and Outdoor set-up. However, the framework couldn't perform out a quantitative evaluation of lack because of absence of ground truth information, un-textured objects like white walls, regular patterns and reflective surfaces became a problem for the system architecture in an outdoor environment. The experimental outcome demonstrated their proposed system was fit for exploring in a pre-arranged paths condition emphatically. However, the data association on the mobile robot turned into a difficult because of varieties in light, specular appearance in environment and conflicting point clouds which happened as a result of the variety in review edges of the camera.

Visual Simultaneous Localization and Mapping (VSLAM) has been a common method among robotic applications because of its ability to enable an autonomous robot to construct a model of its environment and estimate its position at the same time with the aid of vision sensor such as camera etc. This is an important functionality necessary for a robot to navigate in an unfamiliar area. Numerous researchers have researched in an area of Visual Simultaneous Localization and mapping (VSLAM) with remarkable results. However the review conducted in this study has discovered that challenges such as variations in illumination, specular reflection in environment, inconsistent point clouds, un-textured objects like white walls, regular patterns, reflective surfaces, direct sunlight or

shadows illumination, loop closure detection and multiple camera sensors on the robot has the potential to limit or cause the Visual Simultaneous Localization and Mapping (VSLAM) technique to fail whether in a static or dynamic environment. Thus, in an attempt to propose a better SLAM technique that is camera based, this study will be attempting to solve the problem of light intensity in a static environment work because this is often common as a result of high intensity of sunlight in our continent by introducing algorithm into the system that will overcome these problems mentioned in our literature review. Some of these algorithms are mentioned below.

B. Algorithm for Localization and Mapping

The capability of an autonomous machine to construct a model of its environment and to estimate its position is an important functionality necessary for a robot to navigate in an unfamiliar area [18]. Present robotic systems such as SLAM need the exact relationship between the forecast and detected features, for updating the robot's position, and model. The incorrect data relationship between the forecast and detected features will cause an unreliable model and erroneous localization [19]. In an attempt to address the SLAM issue of environmental noise such as light intensity in our system, we assume that a robot is navigating a static environment where stationary landmarks such as trees and building exist; a study will be conducted and illustration of an algorithms that can be used to addressing static landmark in an environment is given below.

According to [12] the original SLAM posterior is given as follows:

$$p(x_{1:t}, l_{1:m} | z_{1:t}, u_{o:t-1}) \quad (1)$$

where $x_{1:t}$ represents the path of the robot, $l_{1:m}$ represents the landmarks,

$z_{1:t}$ represents the measurements and

$u_{o:t-1}$ represents the commands given to the robot.

As for static landmarks, the landmark will be defined as follows:

$$l_{1:m} = \{ l^{s}_{1:m_s} \} \quad (2)$$

where m represents the number of landmarks, static landmarks. Static landmarks are separated from static as follows. Then (1) is divided into two parts as follows:

$$p(x_{1:t}, l_{1:m} | z_{1:t}, u_{o:t-1})$$

$$= p(x_{1:t}, l^{s}_{1:m_s} | z_{1:t}, u_{o:t-1}).$$

$$= p(x_{1:t}, l^{s}_{1:m_s} | z_{1:t}, u_{o:t-1}) \quad (3)$$

Also, static landmarks can be independent from each other as it's shown below:

$$p(x_{1:t}, l^{s}_{1:m_s} | z_{1:t}, u_{o:t-1})$$

$$= p(x_{1:t}, l^{s}_{1:m_s} | z_{1:t}, u_{o:t-1}). \quad (4)$$

C. Image filtering Algorithms

Filtering algorithms are vital method in image processing, they are used to decrease noises in image. Environmental noise has the competency to destroy image, corrupt the RGB color value and bring poor vision that makes the image content interpretation difficult to analyses [19]. According to [19] they are several types of Environmental noises (snow, Shadow, fog, humility, rain, dew) exist, but in this research out attention is only focused on light intensity because of its common occurrence on a daily basis. Below is the illustration of a light intensity filtering algorithm

Light intensity algorithm has the capability to minimize the effect of light intensity affecting the image [19]. The technique is first carried out on modelling the object reflected by camera based on dichromatic reflection represented as I(x).

Thus, the light intensity detection method is based on the use of dark channel shown in equation (5) and automatic thresholding illustrated in equation (6) is used to label high light reflection in image. These two technique is used for identification for high light area because the intensity value of such are in the dark channel will be having high intensity value while non-area affected with light will be having low intensity value.

I^{dark} of **I** can defined as follows:

$$I^{dark}(X) = \min_{y \in \Omega(x)} (\min_{c \in \{r,g,b\}} (I^{c}(y))) \quad (5)$$

$\Omega(X)$ is a local patch cantered at x,x = {x,y} is the image coordinates, I^{c} is a color channel of ***I***

$$t^{*} = ArgMax\{(1-p_t)(\omega_1(t)\mu_1^{2}(t) + \omega_2(t)\mu_2^{2}(t))\} \quad (6)$$

Where *t* is a threshold value, p is the probability of occurrences at threshold value *t,* the smaller the P*t* is, and the larger the weight will be.

In the mark image generated by automatic thresholding of dark channel image, the area labelled as 1 signifies the area affected with light intensity and 0 signifies area that are not affected with light intensity. Illustration for these expression is given in equation (9).

$$masmask(x) = \{ {1 \atop 0} \; if \; I(x) > t^{*} \quad (7)$$

In the removal process of Light intensity, Specular to diffuse mechanism is proposed in this technique, illustration of the image with minimal light intensity effect I^{D} is given in equation (10).

$$I^{D}(\wedge_{\max}) = I - \frac{\max_{u \in (r,g,b)} I_u - \wedge_{\max} \sum_{u \in (r,g,b)} I_u}{1 - 3\wedge_{\max}} \quad (8)$$

III. PROPOSED VISUAL SIMULTANEOUS LOCALIZATION AND MAPPING FRAMEWORK

This section gives an outline clarification of our Vision based Simultaneous Localization and Mapping (VSLAM) technique for the independent robot effectively manufacture a map and discover its position inside the location. Figure 1 gives a graphical illustration of our framework and followed by a discussion of each phase in our Visual Simultaneous Localization and Mapping (VSLAM) technique

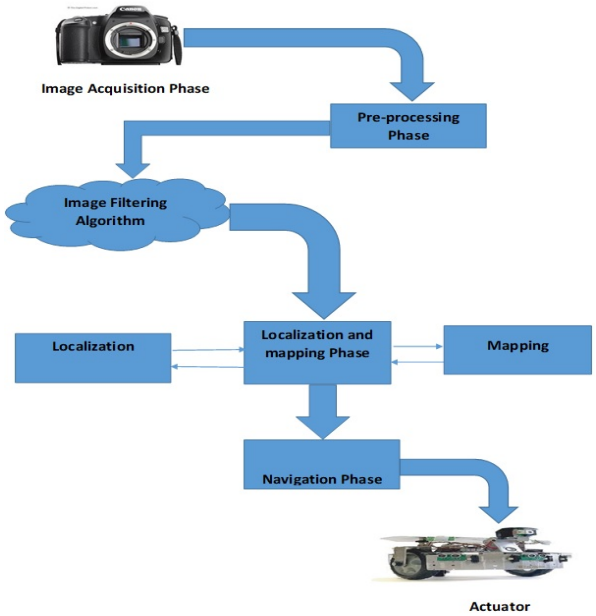


Fig 1: Proposed Framework of Visual Simultaneous Localization and Mapping

IV. FRAMEWORK DISCUSSION

A. *Sensors (acquisition stage)*

Before image processing can start, a picture must be received by a sensor and transformed into a controllable object; this method is called image acquisition [20]. According to [20] the image acquisition procedure involves three phases: energy reflected from the entity of concern, an optical system which emphasizes the energy and a sensor which calculates the quantity of energy. The energy of interest is normally electromagnetic waves [20] and the light reflected from the object has to be taken by the sensor. If a material sensitive to the reflected light is located near to the entity, a picture of the entity will be taken [20]. The light reflected from the object of concern is focused by some optics and now must be logged by the sensor [20]. The image will then be sent to pre-processing stage for further processing.

B. *Pre-process Algorithm*

The second phase in our method is Image pre-processing, since the major part of the genuine information from camera sensor are noisy, incomplete and inadequate, so pre-processing from acquisition stage becomes essential [21]. Image pre-processing was one of the preparatory stages which were very required to guarantee the high exactness of the subsequent stage [21]. [22] defined Image Pre-processing as a procedure to enhance crude pictures caught from cameras/sensors situated on satellites, space tests and air ships or pictures caught in standard day by day life for various frameworks. The reason for pre-pre-processing was an upgrade of the picture data that overwhelms unfavorable misrepresentations or enhances some picture features pertinent for extra processing with and analysis work [23]. The photos taken from the vision sensors experience pre-process stage to be prepared, analyzed for an autonomous robot to extract the landmarks on the picture. In our work to improve Visual Simultaneous Localization and Mapping (VSLAM), at the pre-processing stage, we had introduced a noise filtering algorithm to remove the noise on pictures which were harmed by higher light intensity in the Acquisition stage. The next section introduces the filtering algorithm to remove the environmental noise such as light intensity had been removed from acquired image.

C. *Image Filtering Algorithm*

An Image filtering algorithm is vital to our technique as it is utilized to lessen environmental noise, clean and improve the pictures that are as of now influenced by environmental noise, for example, light intensity, shadow, rain and so forth. Environmental noise, for example, light variety are common noise most particularly in office territories [8] and can cause the Visual Simultaneous Localization and Mapping (VSLAM) method to fail, because as mentioned early in the study the environmental noise can harm picture, degenerate the Red Green Blue (RGB) shading value and bring decreased vision [19]. As mentioned early, in our pre-processing stage we had introduced a noise filtering algorithm, but in this research our attention is only focused on light intensity because of its common occurrence on a daily basis. After filtered images with uncorrupted features generated will now be sent to localization and mapping for further processing of the image which will lead to minimum error in the estimation of the environment [19].

D. *Localization and mapping*

The third phase of our technique localization and mapping, the whole procedure and motivation behind simultaneous Localization and Mapping (SLAM) is develop a model of its location and estimate the robot position inside location simultaneously [8]. This can't be accomplished until the point that robot can refresh its odometry which gives guidance to the robot [24]. Odometry it is continually deceptive in begin of the robot navigation, so the robot needs to explore location in order for vision sensor as an exteroceptive sensor to obtain the scene of the environment in order to redress the robot position [24]. The vision sensor gives scene of the location in form of a picture which needs to go pre-processing phase to filter environmental noise acquired on acquisition phase [24]. The robot would extricate the features from the obtained picture of the earth and re-observing the scene the robot returns to the territory [25].

E. *Navigation Algorithm*

Another fundamental phase in our proposed approach is Navigation phase which is responsible for robot movement around the environment. Navigation is a science of getting mobile robots from place to place [26], the ability of a mobile robot to navigate without external aid [27]. The autonomous navigation of robots in uncontrolled

environments is a challenge because it requires a set of subsystems to work together. It requires building a map of the environment, localizing the robot in that map, making a motion plan according to the map, executing that plan with a controller, and other tasks; all at the same time [28]. According to [29] a robot application utilizes path planning and local motion controls to navigate the unknown territory. Path planning studies a model or a map of the location to decide on the regular path points for an autonomous robot to trace from a start site to its destination. Local motion utilizes sensory data to decide a movement that will evade crash with unknown objects or objects whose station in the location had changed [29]. After path planning and local motion, the navigation algorithm will send a command to actuator to instruct the movement of the robot within the unknown location. There are many navigation algorithms which are available from literature each with its strengths and drawbacks however in our method, A* Algorithm is adopted this study to plan a local optimal collision-free path from the current location of the robot to because of its ability to reduces the number of node explorations with respect to Breadth- and Depth-First [28].

V. Proposed Evaluation Approach

This study intends to investigate the performance of the proposed Visual Simultaneous Localization and Mapping (VSLAM) framework by performing experiments in the following ways:

- Vision Sensor in a form of a camera will be utilised to acquire data and information from the environment.
- The Internet will be used as an additional source to acquire data and information.

The experimental work and simulation will be executed using Matlab Software. The experimental work will be quantitatively evaluated by the three ratios below, and qualitatively evaluated by comparing with existing results from the literature to reveal the system with the best performance.

- *The false positive ratio (FPR*): Represents the proportion of paths case that are erroneously classified as non-road.
- *The false negative ratio (FNR):* The proportion of paths case that are properly classified.
- *The accuracy:* The proportion of a total number of paths and non-roads pixels that are classified properly.

VI. Conclusion

In this study we have proposed a framework for Vision Based simultaneous localization and mapping (VSLAM) framework, to enable autonomous robot correctly to navigate the environment. We also adopted the filtering algorithm in our VSLAM framework to enable it to reduce or limit the amount of light intensity on images taken from the environment when an autonomous robot navigates the area. Images acquired from the environment, go through the process of pre-processing where filtering algorithm has been introduced to filter the noise cause by the environmental factors such light, rain, snow etc. This process gives VSLAM system the ability to correctly identify the features or landmarks on the environment, from identified features or landmarks, from identified landmarks the VSLAM system associates currently mapped landmarks with newly mapped landmarks and update the odometry data for the autonomous robot to successfully localizes itself and builds the map of environment.

References

[1] G. Narang, K. Nakamura, and K. Nakadai, "Auditory-aware navigation for mobile robots based on reflection-robust sound source localization and visual slam," pp. 4021–4026, 2014.

[2] Y. Zhang, T. Zhang, and S. Huang, "Comparison of ekf based slam and optimization-based slam algorithms," pp. 1308–1313, 2018.

[3] S. Khan, D. Wollherr, and M. Buss, "Modeling laser intensities for simultaneous localization and mapping," IEEE Robotics and Automation Letters, vol. 1, no. 2, pp. 692–699, 2016.

[4] L. Xiang, Z. Ren, M. Ni, and O. C. Jenkins, "Robust graph slam in dynamic environments with moving landmarks," in Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on. IEEE, 2015, pp. 2543–2549.

[5] J. O. Esparza-Jimenez, M. Devy, and J.-L. Gordillo, "Ekf-based slam fusing heterogeneous landmarks," in 17th International Conference on Information Fusion (FUSION), 2014.

[6] W. Fang, Y. Zhang, B. Yu, and S. Liu, "Fpga-based orb feature extraction for real-time visual slam," in Field Programmable Technology (ICFPT), 2017 International Conference on. IEEE, 2017, pp. 275–278.

[7] C. Park and J.-B. Song, "Illumination change compensation and extraction of corner feature orientation for upward-looking camera-based slam," in Ubiquitous Robots and Ambient Intelligence (URAI), 2015 12th International Conference on. IEEE, 2015, pp. 224–227.

[8] X. Gao and T. Zhang, "Robust rgb-d simultaneous localization and map-ping using planar point features," Robotics and Autonomous Systems, vol. 72, pp. 1–14, 2015.

[9] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendon´-Mancha, "Visual simultaneous localization and mapping: a survey," Artificial Intelligence Review, vol. 43, no. 1, pp. 55–81, 2015.

[10] C. Park and J.-B. Song, "Illumination change compensation and extraction of corner feature orientation for upward-looking camera-based slam," in Ubiquitous Robots and Ambient Intelligence (URAI), 2015 12th International Conference on. IEEE, 2015, pp. 224–227.

[11] A. Yassine, C. Fatima, and B. Aissa, "Monocular slam based on salient regions detection," in Electrical Engineering-Boumerdes (ICEE-B), 2017 5th International Conference on. IEEE, 2017, pp. 16.

[12] S. Oh, M. Hahn, and J. Kim, "Simultaneous localization and mapping for mobile robots in dynamic environments," in Information Science and Applications (ICISA), 2013 International Conference on. IEEE, 2013, pp. 1–4.

[13] S. Maity, A. Saha, and B. Bhowmick, "Edge slam: Edge points based monocular visual slam." in ICCV Workshops, 2017, pp. 2408–2417

[14] Y. Abdelrasoul, A. B. S. H. Saman, and P. Sebastian, "A quantitative study of tuning ros gmapping parameters and their effect on performing indoor 2d slam," in Robotics and Manufacturing Automation (ROMA), 2016 2nd IEEE International Symposium on. IEEE, 2016, pp. 1–6.

[15] K. N. Al-Mutib, E. A. Mattar, M. M. Alsulaiman, and H. Ramdane, "Stereo vision slam based indoor autonomous mobile robot navigation," in Robotics and Biomimetics (ROBIO), 2014 IEEE International Con-ference on. IEEE, 2014, pp. 1584–1589.

[16] R. Raguram, O. Chum, M. Pollefeys, J. Matas, and J.-M. Frahm, "Usac: a universal framework for random sample consensus." IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 2022–2038, 2013.

[17] C. Brand, M. J. Schuster, H. Hirschmuller, and M. Suppa, "Stereo-vision based obstacle mapping for indoor/outdoor slam," in Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on. IEEE, 2014, pp. 1846–1853.

[18] G. Narang, K. Nakamura, and K. Nakadai, "Auditory-aware navigation for mobile robots based on reflection-robust sound source localization and visual slam," in Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference on. IEEE, 2014, pp. 4021–4026.

[19] A. O. Yinka, S. M. Ngwira, Z. Tranos, and P. S. Sengar, "Performance of drivable path detection system of autonomous robots in rain and snow scenario," in Signal Processing and Integrated

Networks (SPIN), 2014 International Conference on. IEEE, 2014, pp. 679–684.

[20] T. B. Moeslund, Introduction to video and image processing: Building real systems and applications. Springer Science & Business Media, 2012.

[21] N. Shameena and R. Jabbar, "A study of preprocessing and segmentation techniques on cardiac medical images," 2014.

[22] A. Ballabeni, F. Apollonio, M. Gaiani, and F. Remondino, "Advances in image pre-processing to improve automated 3d reconstruction." International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences, 2015.

[23] M. Sonka, V. Hlavac, and R. Boyle, Image processing, analysis, and machine vision. Cengage Learning, 2014.

[24] H. Williams, "Human inspired robotic path planning and heterogeneous robotic mapping," 2016.

[25] F. Heukels, "Simultaneous localization and mapping (slam): towards an autonomous search and rescue aiding drone," Master's thesis, University of Twente, 2015.

[26] U. Frese and G. Hirzinger, "Simultaneous localization and mapping-a discussion," in Proceedings of the IJCAI Workshop on Reasoning with Uncertainty in Robotics. Seattle, 2001, pp. 17–26.

[27] G. Kosuru, "Design and implementation of an ekf based slam algorithm on a mobile robot," Ph.D. dissertation, International Institute of Infor-mation Technology Hyderabad-500 032, India, 2011.

[28] E. Z. Gomez, "Map-building and planning for autonomous navigation of a mobile robot," Ph.D. dissertation, Center for Research and Advanced Studies of the National Polytechnic Institute, 2015.

[29] A. Bircher, M. Kamel, K. Alexis, M. Burri, P. Oettershagen, S. Omari, T. Mantel, and R. Siegwart, "Three-dimensional coverage path planning via viewpoint resampling and tour optimization for aerial robots," Autonomous Robots, vol. 40, no. 6, pp. 1059–1078, 2016.

Employment Law Expert System

Preetila Seeam, Nishant Teckchandani, Hansha Booneyad and V. P. Torul
Aberystwyth University (Mauritius)
Uniciti, Flic en Flac
Mauritius
Email: {prs12,nit23,jdb19,vpt} at aber.ac.uk

Amar Seeam
Middlesex University (Mauritius)
Uniciti, Flic en Flac
Mauritius
Email: {a.seeam} at mdx.ac.mu

***Abstract*—This paper reviews the development and application architecture of an expert system to assist the Mauritian population with queries they may have about labor or employment law. The expert system makes use of Machine Learning, Speech Recognition/Synthesis and Natural Language Processing techniques to converse with users through a web interface. The expert system also takes advantage of a large knowledge base, that allows the system to teach itself employment law principles. The knowledge base is created from "Understanding Employment Law and Remuneration Orders in Mauritius", written by Ved Prakash Torul [1], which is a simplified version of the Employment Relations Act and the Employment Rights Act. The book explains employment law in common language, to help the public understand their constitutional rights. The expert system allows users to communicate and express their employment issues, so that they are aware of their next course of action, either they are an employer, employee, or a union. The paper also reviews the evaluation period, which consisted of a preliminary testing period. Through the evaluation, it was concluded that the expert system was able to respond to individual responses with a Precision of 66% and Recall of 85%. While the Expert System is able to converse with users on certain topics on Employment Law, further evaluation would need to be conducted. Additionally, the knowledge base will need to be updated over time.**

***Index Terms*—Expert System, Deep Neural Network, Knowledge Base, Employment Law, Web Interface**

I. INTRODUCTION

The Employment Relation Act 2008 and the Employment Rights Act 2008 opened a new plethora of rights and duties for the employers, employees and the Trade Unions in Mauritius. This development in employment law has brought an enormous shift in the employer and employee relationship, which was initially founded on power and subordination. The emergence of this economic commodity lead to an increase in productivity, profit for the enterprise, social dialogue and active participation in decision making. Accessibility to information on employment law is crucial when the complexities emerging through the pressures to establish a conflict free workplace and to create a meaningful relationship between the employer and employee exist. It requires every employer and employee, trade union, human resource manager, student, legal practitioner and even lay people to keep abreast with substantive, as well as procedural aspects of employment law. It is noted that the modern approach of dealing with employment issues is no longer the domain of lawyers only. The Employment Relation Act and the Employment Rights Act have introduced principles of employment law that affect many players in the Mauritian economy. It has made it increasingly important for all those functioning in the commercial world to have a sound understanding of the innovative provisions of employment law and the industrial relations. With this in mind Professor Ved Torul, who is an expert in these matters, has created a lengthy manual covering these issues [1].

The manual encompasses the following :

1) Contract of Employment
2) Statutory rights of employees and workers.
3) Dismissal for misconduct, poor work performance and operational requirements.
4) Trade unions and organizational rights.
5) Dispute resolution procedures.

A tremendous amount of knowledge and information is contained within this manual - however retrieval of a particular issue still requires an understanding of the law and a deep appreciation of what to look for. Therefore it's accessibility is limited, where it needs to be available to all, including those who have low levels of literacy (i.e. cannot read nor write). Taking all this into account, the main aim of the research is to computerize this manual and to create a queryable employee law knowledge based system with an expert system component. This component contains elements of artificial intelligence that can help and guide the employee towards finding a solution to any issues they may have.

II. BACKGROUND

Expert Systems are a branch of Artificial Intelligence (AI). AI is primarily concerned with the development of computer systems that exhibit reasoning characteristics associated with human intelligence. AI also includes pattern recognition, machine learning and natural language. Heuristics add additional reasoning understanding and judgment (decision making).

Expert Systems are designed to perform tasks carried out by human experts with the ability to help aid in decision making processes using a knowledge base developed by an expert or a representation of the decision making process by an expert.

Knowledge Based Systems are those that are concerned with the storage of knowledge and the data structures whereas an Expert System has the ability to act on the knowledge and further process it according to an experts reasoning capabilities this can be achieved using either Rule or Case based techniques.

An expert system can be used to make choices that would normally be performed by a domain expert. Those seeking the knowledge of an expert, can make use of the wide range of collective knowledge, and Artificial Intelligence techniques used by an Expert System. The expert system can be used to support individual employee cases - i.e. a Case based expert system. Expert Systems can in fact be categorized further into two types of system - Rule and Case based - and we are proposing to implement both.

A. Expert Systems

Expert Systems have come along away since MYCIN origins [2], in the 1970s. They have seen successful implementations in the legal [3], medical [4] and law enforcement [5] fields and enable expert human knowledge to be stored, refined, processed and retrieved through the use of computer information systems. [6] critically reviewed numerous systems, noting in particular the challenges faced for legal expert systems, such as having adequate reasoning. In saying this, we still believe that a system is needed for the Mauritian context, particularly one with speech synthesis to aid ease of use. Though Mauritius has seen rapid development in the computerization of the legal system through the e-Judiciary systems, its use is exclusively for the lawyers, law enforcement agencies and the courts. One recent example of the use of expert system in law in Mauritius was carried out by S. Pudaruth et al. [7], who developed the system to to ascertain whether two individuals can legally be married according to Mauritian family laws. Similarly a question answer system for the Mauritian judiciary was described in the other paper by S.Pudaruth et al [8].

The Mauritian judiciary system detects keywords in sentences, and produces responses, in the form of links to cases, for the specific topics mentioned. This keyword detection is similar to AIML (Artificial Intelligence Mark-up Language) [9], or a search engine query. The expert system discussed in this paper tries to emulate a conversation with a lawyer, and provides solutions to personalized problems users may have. The Employment Law expert system also makes use of context in a conversation, and patterns in sentences to detect what kind of response the user is looking for. Additionally, speech recognition and synthesis is used for adding multiple dimensions of conversation methods.

B. Expert System Overview

Fig. 1 shows the various components of an expert system.

1) Knowledge Base: The Knowledge Base contains all the rules, cases, facts and information.

2) Inference Engine : This component is the deductive reasoner that will process input according to the knowledge base (i.e. rules, cases, etc).

3) User Interface : The user interface permits interaction with the system - modern implementations are web-based.

4) Explanation Function: This components determines how to present the information to the user

5) Knowledge Builder: This component is the interface for the expert or someone who will be entering expert knowledge into the system.

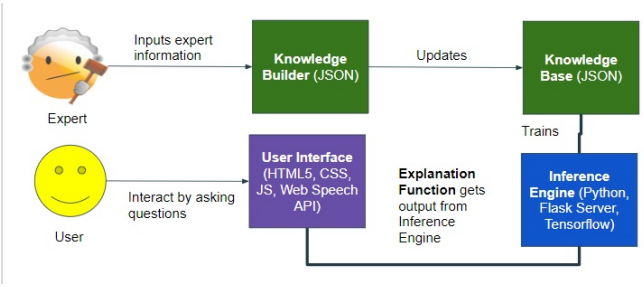


Fig. 1. Expert System Architecture

C. Rule Based

Some computer programs utilise an orthodox approach in problem-solving by making use of structured algorithms and data structures in discovering solutions, while conventional rule-based expert systems utilise human intelligence; knowledge exhibited by experts are most times based on rules or data that is stored in the computer. Thus for problems to be resolved, rules and data are required by the experts. For example, the type of expert system used to provide legal advice is known as a rule-based legal expert system.

D. Case Based

Case-based reasoning frameworks are a method of demonstrating knowledge application through explicit historic cases. This methodology is not the same as the rule-based based approach in light of the fact that the knowledge is not ordered and deciphered by an expert; rather, the experiences that formed the experts knowledge are specifically utilised as a part of making decisions. The issue with the case-based reasoning is learning, in light of the fact that with the insignificant expansion of new cases to information repository. Regardless, to resolve issues using this system a substantial level and size of knowledge repository must be provided to support the related problems.

E. Expert Systems related to Employees

A literature review has indicated that no equivalent system exists that aids employees to navigate through the complexities of employee law to aid them in decision making or gather necessary information. Related applications include those by [10] who investigated Employee skills analysis using a hybrid neural network and expert system; [11] demonstrated an Employee Benefit System, using expert systems; [12] developed an employee involvement model based on expert systems; [13] created an expert system for screening employee pension plans for the Internal Revenue Service.

F. Evaluation Methods

Numerous methods have been used to evaluate the performance of expert systems, especially with conversational agents, like the Employment Law Expert System. Based on the properties and aim of the system, different evaluation methods will have to be used [14]. User satisfaction scores, or task completion tests were the common methods of evaluation [15], however, evaluation types are still experimental when dealing with intelligent conversational agents, or systems using end-to-end training with deep learning [16]. One of the first methods of evaluating a chatbot or conversational model was the Turing test [17]. This was a study that replaced a human with a machine to deploy a certain task like reading out text or create an image. A human observer would then analyze the output and determine if the task was completed by a machine or a human. However, studies have shown that the Turing test is not an effective evaluation method of determining the performance of an expert system, and is considered biased based on the observer evaluating the system [18]. Algorithmic methods of evaluation have been studied like the BLEU Score [19]. BLEU uses the dataset, for the purpose of training the system, as a reference of what a good quality response is. It then calculates a score on each response by comparing the sequence given by the user to the good quality response discussed earlier [19]. An average score will be calculated based on all the responses from the conversation. BLEU gives outputs from 0 to 1 where 1 is the most accurate, while 0 represents a wrong response. BLEU is a cheap and effective way of evaluating an intelligent conversational system, due to its lack of testers. Finally, Information Retrieval can be used to find how a conversational expert system is performing per response [20]. This includes analyzing each response to a user's input sequence. Each response can be given a correct or incorrect response indication, and that would have been enough. However, there are user questions that grammatically do not make sense, spelled incorrectly, or just contain a word. Precision and recall take into consideration the expert system's ability to detect correct or incorrect responses, to both the users correct and incorrect questions.

III. Scope of Work

The outcome of the research will be to create an expert system which will be used by employers, employees, Human Resource Managers, General Managers, students and lay persons. The benefits of this system are as follows:

1) To help the employees and employers to gain access to pertinent issues on Employment Law and industrial trial relations through a computerised system.
2) To enable employees and employers to grasp complex employment issues with least technicalities. They will feel empowered to exercise their rights, to assist in dealing effectively with misconduct, incapacity due to ill health and poor work performance. The expert system will act as a guide for them.
3) To understand the salient points regarding substantive and procedural fairness during disciplinary hearing. The expert system's knowledge base will build and monitor individual cases.
4) To set up and execute a grievance procedure that is legal, fair and proportionate. The expert system will only draw up about the laws contained within it's knowledge base to ensure that these scenarios are executed within the boundaries of employee rights.
5) It is hoped that in the future the system can be linked up and made inter-operable with various relevant websites, including the Supreme Court, Ministry of Labour, Industrial Relations, Employment and Training, Commission for Conciliation and Mediation, amongst others.

IV. Implementation of Expert System

A. Knowledge Base and Knowledge Builder

Firstly, the knowledge base was set up so that the data entered by the expert could exist in an organized manner. This data would later run through the inference engine. The data is saved into a JavaScript Object Notation (JSON) file, with each block of data called an Intent. This is similar to the Language Understanding Intelligent Service (LUIS) dialogue system that Microsoft makes use of [21]. However, to analyze the context of the conversation more clearly, each intent contains the intent name (tag), possible patterns the user may type (patterns), possible responses (responses), and a context identifier. Fig. 2 displays 2 examples of intents in the knowledge base.


```
{"tag": "misconduct",
"patterns": ["What qualifies as misconduct?", "What is considered misconduct", "How did I misconduct?"],
"responses": ["Misconduct includes the following:<ul><li>Breach of trust and confidentiality</li><li>Dishonest behaviors</li><li>Una
"context_set": "misconduct"
},

{"tag": "misconductBreachOfTrust",
"patterns": ["What is a Breach of Trust?", "What could be confidential for a company?"],
"responses": ["Breach of Trust and confidentiality includes the following: <ul><li>Fraud</li><li>Theft</li><li>Unauthorised posessio
"context_filter": "misconduct"
},
```


Fig. 2. Example of Knowledge Base JSON structure

With the JSON file, it is possible to organize documents (sentences), words, and classification classes (intents) as shown in Fig. 3 .


```
131 documents

34 classes ['Employee', 'Employer', 'TORUL', 'arrested', 'arrestedQuery', 'contractAgreement',
'expressTerms', 'goodbye', 'greeting', 'lengthOfContract', 'lengthOfNotice', 'misconduct', 'mi
'misconductLateness', 'misconductViolence', 'name', 'reasonsForDismissal', 'renumerationExampl
'renumerationExampleNoPaymentMonthlyContract', 'renumerationExampleNoPaymentNoLabourOffice', '
'statementOfParticulars', 'statementOfParticularsSending', 'terminationContractDefinition', 't
'terminationContractProcedure', 'thanks']

166 unique stemmed words ['!', "'s", ',', '.', '2', 'a', 'accord', 'aft', 'agr', 'allow', 'am'
'at', 'back', 'be', 'been', 'bef', 'behavio', 'between', 'boss', 'breach', 'by', 'bye', 'cal',
'day', 'deem', 'defin', 'definit', 'detain', 'determin', 'did', 'dishonest', 'dismiss', 'do',
'get', 'giv', 'go', 'good', 'goodby', 'govern', 'has', 'hav', 'he', 'hello', 'help', 'her', 'h
'labo', 'lat', 'law', 'left', 'leng', 'limit', 'long', 'many', 'mauriti', 'me', 'mean', 'meant
'nee', 'not', 'oblig', 'of', 'off', 'on', 'or', 'oth', 'own', 'paid', 'particul', 'partiul', '
'ref', 'refus', 'reg', 'rel', 'renum', 'requir', 'resign', 'right', 'say', 'see', 'serv', 'sha
'termin', 'thank', 'that', 'the', 'ther', 'thi', 'tim', 'to', 'tor', 'trust', 'two', 'viol', '
'writ', 'yo', 'you']
```


Fig. 3. Prior to training the model, a list of documents (131 sentences), classes (34 intents), and a list of stemmed words (166 unique words) are created

B. Inference Engine

Python and Tensorflow were used to write the software for the inference engine. Tensorflow is an open source software library to assist developers to use Machine Learning algorithms [22]. In the case of the inference engine, a Deep Neural Network (DNN) was used. The DNN trains the inference engine to identify the specific intent, when a user types in a query [23]. The DNN consists of an input layer, 2 hidden layers, an output layer, and a regression algorithm using gradient descent. The input layer consists of 5 features (each component consisting in an intent from the JSON file), the hidden layers consist of 8 nodes, and the output layer consists of 2 classes (the output intent and context of the conversation) as seen on Fig. 4. The regression includes a gradient descent function (sigmoid function), to normalize values and its derivatives to measure the error rate (loss value). The training keeps iterating the values to keep the error rate as low as possible.

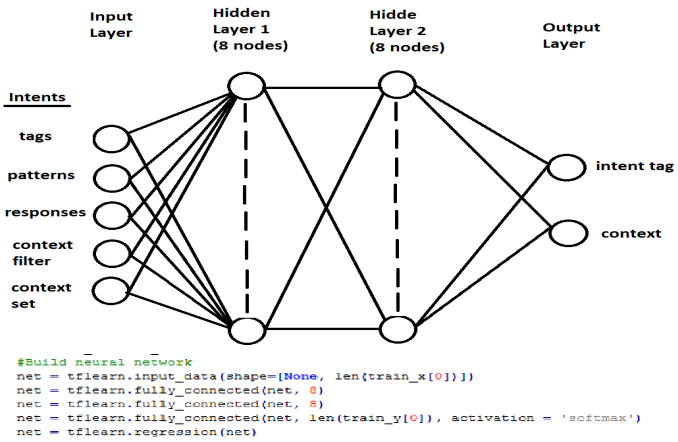


Fig. 4. The architecture of the DNN consisting of an input layer, 2 hidden layers, and an output layer. The input takes in the intents with the corresponding categories in the intent, and returns the context and responses at the output layer

The DNN is not able to pass words and sequences into its network, therefore, the words were converted to tokens of arrays (integer values). In order to do so, the Python Natural Language Tool Kit (NLTK) [24] was used. The NLTK converted the data from the Knowledge Base and passed it through the DNN. After training, Tensorflow allows developers to view their values calculated via their Tensorboard data visualization feature. For the DNN used in this inference engine, the Loss and Accuracy were calculated throughout the training period. The Loss value is a summation of errors made per training data point (intent), and the Accuracy shows the percentage of correct outputs the DNN generates. As seen in Fig. 5, the Loss moves closer to 0 over time, which displays that the inference engine made fewer errors as training went on. The Accuracy also approached 1.0.

After training the model, classifying the correct intent can be tested with the user's input message. When the model receives an input message, the DNN model will return the

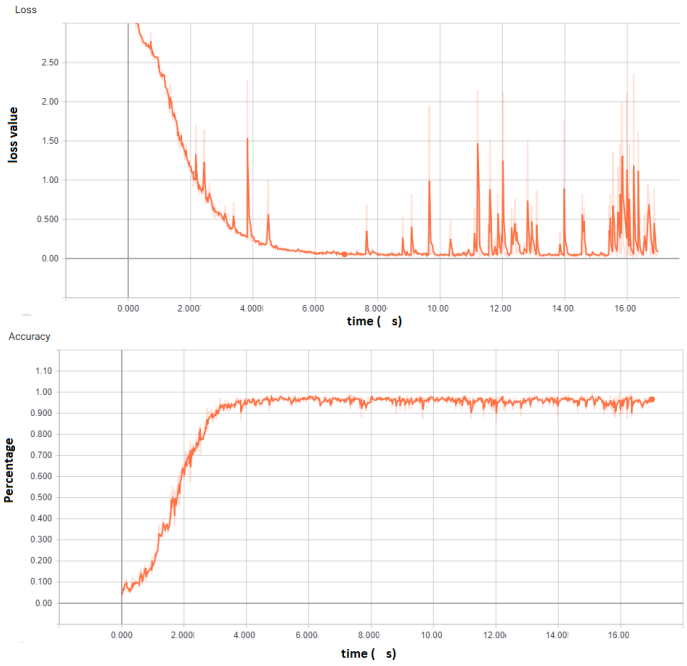


Fig. 5. The Loss (top) and Accuracy (bottom) of the DNN over training time

intent and context according to the training it received. This is displayed in Fig. 6.

```
response(What is a contract of employment?)

<p>A contract of employment is an agreement in terms of which the
employee/worker agrees to make his personal services available to the employer
under the latter's supervision and authority in return for remuneration.</p><p>
A contract of employment may also be a contract of service or of apprenticeship
and may be oral, written, implied or express.</p>

classify(What is a contract of employment?)

[('employmentContract', 0.9969528)]
{'123': 'employmentContractDefinition'}
```

Fig. 6. The response function takes in the user's input message, and returns a response message. The classify function returns the intent tag and its corresponding probability, plus the context of the conversation below it.

C. User Interface and Explanation Function

A web interface provided a platform for the user to interact with the system, to get their queries answered. The application was run on a FLASK server, which is written in Python. A FLASK server was used, so that appropriate interaction with the inference engine, which is also written in Python, could be achieved. HTML5, CSS3, and JavaScript were implemented to create the front-end application, and a REST API web service was built so that user queries, and answers generated from the inference engine could be acquired from the FLASK back-end. Addtionally, the Web Speech API [25] was used to activate Speech Recognition capabilities of the user's queries, and Speech Synthesis for the output audio generated by the expert system. Fig. 6 displays the web interface being used with a sample conversation by a user.

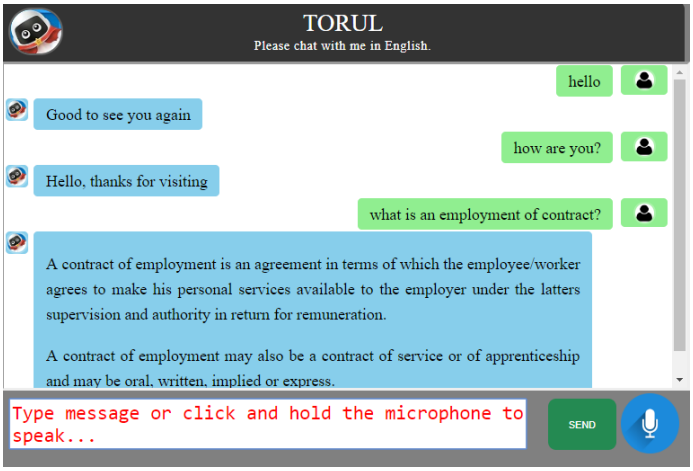


Fig. 7. An example of the web interface being used

V. Experimental Setup

After the development of the Employment Law Expert System, the final stage was to implement an evaluation of the system itself. Information Retrieval rate was used to evaluate the Employment Law Expert System. To proceed with the evaluation, 5 university staff members participated in testing the Employment Law Expert System. A brief on the study, along with seeking ethical consent from the participants was displayed. Each participant was asked to communicate with the system, and ask questions on the topics of contract termination, renumeration, and basic employment law definitions, through text or speech. After each member finished communicating with the expert system, the user's input message, and the generated output message were saved into a SQLite database for further analysis.

A. Results

119 responses were generated. Each response was categorized into 1 of 4 necessary types. The 4 types of responses by the expert system are shown in the following list:

1) TP (True Positive) Expert System responded correctly to users correct input message
2) FP (False Positive) Expert System responded incorrectly to users correct input message
3) FN (False Negative) Expert System responds incorrectly to users incorrect input message
4) TN (True Negative) Expert System responds correctly to users incorrect input message

Precision and recall equations are denoted by the equations (1 and 2):

$$Precision = TP/(TP + FP) \tag{1}$$

$$Recall = TP/(TP + FN) \tag{2}$$

The values of TP, FP, FN, TN, Precision, and Recall are displayed in table I and II below.

TABLE I
Information Retrieval Values (Per Response) of the Expert System where the TP, FP, FN, TN are out of 119 responses generated through the testing

	TP	FP	FN	TN
Number of Responses	67	34	12	6

TABLE II
Precision and Recall Values calculated from the Information Retrieval values from Table I

Precision	Recall
0.6634	0.8481

Precision measures the ability of the expert system to answer correct questions inputted by the user [20]. Precision calculated was 0.6634, which shows that the expert system was able to respond correctly to 66.34% of correctly inputted queries. Recall measures the ability to answer correct questions and detect incorrect questions [20]. The recall calculated was 0.8481, which shows the expert system was able to respond to correct and incorrect queries, 84.81% of the time. This is a relatively high value, and shows the expert system can respond to individual user input messages quite accurately. However, in order to truly help people with their employment law queries, a Precision over 0.9 would be needed. The 0.6634 shows that the technology and architecture is usable, and the number will only increase as the knowledge base grows.

VI. Conclusion

The Employment Law Expert System currently involves interactions with the user through text and speech. However, in order to further access the public, the expert system will have to implement many other languages. Mauritians speak the local language of Creole, French, and English. Creole and French communication capabilities will be needed for a more complete system. Additionally, the system should be accessible to the disable, and illiterate. One way of achieving this accessibility is to provide pictures, and videos notifying users of employment issues they may be facing.

Additionally, the evaluation methods will have to include more components to measure its performance. User satisfaction scores, task completion tests, BLEU scores, along with the Information Retrieval rate in this paper will have to be implemented. Additionally, by expanding the knowledge base, and fine tuning the speech feature, the Precision and Recall rate will increase and the system will perform better.

Overall, the development of an Employment Law Expert System, is vital, in order to help the Mauritian population be more aware of their employment rights. The system discussed in this paper, provides an architecture that performs at a very high rate of correct responses. While further work is needed to add features, knowledge, and conduct additional evaluation, the possibilities of this system is endless. The system could be beneficial to the people of Mauritius, and eventually make employment law knowledge more accessible.

ACKNOWLEDGMENT

This work has been funded by the Mauritius Research Council.

REFERENCES

[1] V. Torul and E. Kalula, *Understanding Employment Law and Remuneration Orders in Mauritius*, 1st ed. Juta and Company Ltd.

[2] E. H. Shortliffe, R. Davis, S. G. Axline, B. G. Buchanan, C. C. Green, and S. N. Cohen, "Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system," *Computers and biomedical research*, vol. 8, no. 4, pp. 303–320, 1975.

[3] K. D. Ashley, "Case-based reasoning and its implications for legal expert systems," *Artificial Intelligence and Law*, vol. 1, no. 2-3, pp. 113–208, 1992.

[4] S. Tsumoto, "Automated extraction of medical expert system rules from clinical databases based on rough set theory," *Information sciences*, vol. 112, no. 1, pp. 67–84, 1998.

[5] H. Chen, J. Schroeder, R. V. Hauck, L. Ridgeway, H. Atabakhsh, H. Gupta, C. Boarman, K. Rasmussen, and A. W. Clements, "COPLINK connect: information and knowledge management for law enforcement," *Decision Support Systems*, vol. 34, no. 3, pp. 271–285, 2003.

[6] P. Leith, "The rise and fall of the legal expert system," *International Review of Law, Computers & Technology*, vol. 30, no. 3, pp. 94–106, 2016.

[7] S. Pudaruth and K. Mooneapillay, "An expert system for the mauritian family law," in *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2015, pp. 882–886.

[8] S. Pudaruth, R. Gunputh, K. Soyjaudah, and P. Domun, "A question answer system for the mauritian judiciary," in *2016 3rd International Conference on Soft Computing & Machine Intelligence (ISCMI)*. IEEE, 2016, pp. 201–205.

[9] R. Wallace, "The elements of AIML style," *Alice AI Foundation*, 2003.

[10] F. Labate and L. Medsker, "Employee skills analysis using a hybrid neural network and expert system," in *IEEE International Conference on Developing and Managing Intelligent System Projects, 1993.,*. IEEE, 1993, pp. 205–211.

[11] M. C. Sturman and G. T. Milkovich, "Validating expert systems: A demonstration using personal choice expert, a flexible employee benefit system," *Decision Sciences*, vol. 26, no. 1, pp. 105–118, jan 1995. [Online]. Available: http://dx.doi.org/10.1111/j.1540-5915.1995.tb00839.x

[12] W.-B. Lin, "The exploration of employee involvement model," *Expert Systems with Applications*, vol. 31, no. 1, pp. 69–82, jul 2006. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2005.09.035

[13] G. Grdy and R. S. Patil, "An expert system for screening employee pension plans for the internal revenue service," in *Proceedings of the 1st international conference on Artificial intelligence and law*. ACM, 1987, pp. 137–144.

[14] A. Strigér, "End-to-end trainable chatbot for restaurant recommendations," Master's thesis, KTH Royal Institute of Technology, Sweden, 2017.

[15] S. Möller, R. Englert, K. Engelbrecht, V. Hafner, A. Jameson, A. Oulasvirta, A. Raake, and N. Reithinger, "Memo: towards automatic usability evaluation of spoken dialogue services by user error simulations," in *Ninth International Conference on Spoken Language Processing*, 2006.

[16] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and B. Dolan, "A neural network approach to context-sensitive generation of conversational responses," *arXiv preprint arXiv:1506.06714*, 2015.

[17] A. M. Turing, "Computing machinery and intelligence," in *Parsing the Turing Test*, R. Epstein, G. Roberts, and G. Beber, Eds. Springer, 2009, pp. 23–65.

[18] S. M. Shieber, "Lessons from a restricted turing test," *arXiv preprint cmp-lg/9404002*, 1994.

[19] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[20] K. M. Ting, "Precision and recall," in *Encyclopedia of machine learning*. Springer, 2011, pp. 781–781.

[21] J. D. Williams, E. Kamal, M. Ashour, H. Amr, J. Miller, and G. Zweig, "Fast and easy language understanding for dialog systems with microsoft language understanding intelligent service (luis)," in *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2015, pp. 159–161.

[22] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: a system for large-scale machine learning." in *OSDI*, vol. 16, 2016, pp. 265–283.

[23] I. V. Serban, A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models." in *Proc. 13th Annual AAAI conference*, vol. 16, 2016, pp. 3776–3784.

[24] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009.

[25] J. Adorf, "Web speech api," *KTH Royal Institute of Technology*, 2013.

Hybrid PLC-VLC based on ACO-OFDM

M. D. Kubjana
Department of Electrical and Electronic Engineering Technology
University of Johannesburg
Doornfontein, South Africa
dennis.kubjana@gmail.com

T. Shongwe
Department of Electrical and Electronic Engineering Technology
University of Johannesburg
Doornfontein, South Africa
tshongwe@uj.ac.za

A. R. Ndjiongue
Department of Electrical and Electronic Engineering Science
University of Johannesburg
Auckland Park, South Africa
arrichard@uj.ac.za

***Abstract*—The recent successful development and standardization of the power line communication (PLC) (ITUT G.9960/61, CENELEC EN50065-1, IEEE 1901) and visible light communications (VLC) (IEEE 802.15.7), has resulted in lot of research interest in consideration to the integration of the PLC and VLC technologies. In addition, these technologies present the advantage related to cost effectiveness as a strong research motivation. In this paper, based on the positive aspect of VLC technology, we exploit the Hermitian symmetry to design a hybrid system combining PLC and VLC technologies using orthogonal frequency division multiplexing (OFDM). Asymmetrically clipped optical OFDM (ACO-OFDM) is employed in VLC and investigated when combined with PSK-OFDM (used in PLC). The proposed hybrid PLC-VLC system in this paper takes advantage of the inherent nulling feature of the ACO-OFDM to mitigate impulsive noise that is present in the PLC channel, where the optimal threshold, Tc to combat the impulse is investigated through simulations. While on the other hand taking advantage of the fact that both DCO-OFDM and PSK-OFDM employ the FFT, to achieve a simplified system.**

Keywords—PLC-VLC integration, OFDM, ACO-OFDM

I. INTRODUCTION

The radio frequency (RF) spectrum conventionally used for communication is almost completely depleted, which threatens to leave a significant gap between the wireless access networks and the exponentially increasing wireless data rate demands. Fortunately, the other sections of the electromagnetic (EM) spectrum such as millimetre wave radiation and optical radiation can be used to relieve the RF spectrum [1], as a result both the underlying spectrums are currently under heavy research. The concept of optical wireless communications (OWC) constitutes wireless data communication using the visible light (VL) as a transmission medium. OWC and its subset of visible light communications (VLC) provide several advantages over RF and millimetre wave communications. In addition, the existing lighting infrastructure could be re-used for communication, which facilitates easier system deployment and promises to deliver significant energy savings [2]. The typical light emitting diodes (LEDs) are usually employed due to their long lifetime, relatively small size and energy efficient when comparing with the fluorescent and incandescent light sources. The LEDs illumination possesses a special feature, they promise to transmit data rate at up to 100 Mbps without any regulations on the emissions [3]. The concept of VLC using LEDs for communications has encouraged lighting industries to adapt to LED technology not only for lighting purposes but also for data transmission, the aspect of LED lighting has gained lot of research interests that resulted in the recent development of the IEEE 802.15.7 standard for VLC [4], [5].

While on the other hand, power line communication (PLC) takes advantage of the existing AC electric power transmission infrastructure for data communication, however the PLC networks were not designed with the intention of being used for data communications thus they do not support mobility or the ability to broadcast data over a wide area [2] and the performance is usually limited by inter-symbol interference (ISI), noise and attenuation [6]. The PLC networks are normally used for indoor or local applications such as home automation, monitoring and control. In [7], research has shown that it is indeed possible to achieve high speed communications in this hostile medium with advanced techniques such as orthogonal frequency division multiplexing (OFDM) while overcoming noise and intrinsic multipath inside powerlines.

The recent standardization of PLC (ITUT G.9960/61 [8], CENELEC EN50065-1 [9], [10], IEEE 1901 [11]), has resulted in lot of research interest for the integration of both PLC and VLC technologies. The integration of both PLC and VLC systems presents an attractive solution for low cost communication technology that requires minimal changes to the existing infrastructure while on the other hand sharing the same existing AC electric power cabling. In the literature we looked at various ways already proposed to integrate both PLC and VLC, in amplify and forward scenario using CSK and PSK, [3], [12], [13]. In [6], cascaded PLC-VLC channel using OFDM and CSK techniques is proposed. The system uses OFDM in the PLC channel and the OFDM signal is passed through VLC. The results indicated that the system performance is dependent on the channel. Channel and noise models are proposed in [14], [15] for PLC and for VLC in [16], [17], whereby the PLC channel is influenced by impulsive noise and VLC channel influenced by shot noise.

In this paper, we investigate the effects of impulsive noise generated in the PLC channel and show how it affects the VLC channel which is based on asymmetrically clipped optical OFDM (ACO-OFDM), we also provide solutions that combat impulsive noise at the receiver based on time domain non-linearity techniques such as nulling (blanking) and clipping. The entire PLC-VLC system is then analysed in terms of system bit error rate performance. In ACO-OFDM, the signal is made positive by transmitting only on the odd subcarriers while the even subcarriers are used to form a bias signal and the negative going amplitudes are set to zero (nulling). We therefore take advantage of this inherent nulling mechanism that already exists in the ACO-OFDM, to mitigate the impulsive noise. In addition, the proposed PLC-VLC system is simplified because we make use of the OFDM modulation (and demodulation) only once instead of using it for each channel (PLC and VLC) when compared to already existing systems above.

The rest of the paper is organized as follows, in section II the system model is presented, and brief overview of ACO-OFDM. The performance of ACO-OFDM in PLC-VLC channels with consideration of impulsive noise is investigated through simulations in Section III. Conclusion and remarks are provided in Section IV.

II. System Model

A. ACO-OFDM

In this section a communication model of the proposed technique based on Asymmetrically Clipped Optical OFDM (ACO-OFDM) to realise the integration of PLC and VLC system is depicted in Fig. 1. In this system the contributing noise factor affecting the performance of the transmitted OFDM signal in both the PLC and VLC channels are additive white Gaussian noise (AWGN) and impulsive noise. Throughout the paper the upper cases represent a discrete frequency domain signal and the lower cases represents the time domain signal.

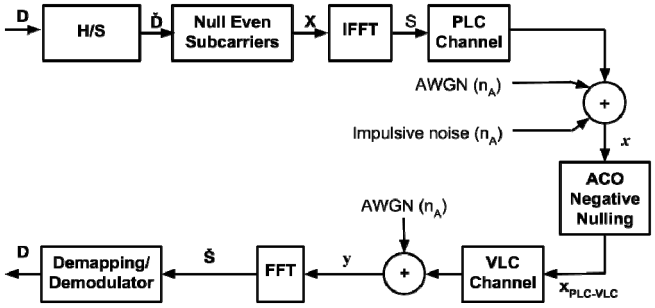


Figure 1. Hybrid PLC-VLC system model based on ACO-OFDM

B. PLC OFDM

OFDM is a special method, scheme or technique for multicarrier modulation which is a derivative of the standard frequency division multiplexing (FDM), obtained by placing closely many orthogonal subcarriers. Since OFDM is a multicarrier system, the transmitted information can be split into smaller multiple chunks independently, therefore allowing for total data rates to be divided into number of subcarriers [18]. The bipolar OFDM signal in the PLC channel consists of N data symbols represented by $D = [D_0, D_1, \ldots, D_{N-1}]$. This OFDM signal is modulated using quadrature amplitude modulation (QAM) or binary phase shift keying (BPSK) grouped in blocks of N parallel symbols. The signal D is passed through an IFFT block in order for conversion to time domain, this signal is now represented by a vector $s(k) = [s_0(k), s_1(k), s_2(k), \ldots, s_{N-1}(k)]$. The digital to analogue converter (DAC) receives the signal and converts appropriately, this time domain signal at the output of the DAC is then transmitted through a channel where it gets corrupted by noise. This paper looks at two types of noise AWGN and impulsive noise modeled as Middleton Class A noise.

C. Impulsive noise modelling

As already mentioned in the previous section Middleton Class A noise model will be used to model the statistical noise for impulsive noise, in practice the Middleton Class A model is based on Poisson-Gaussian model. This model is mostly used in communication systems to model the effects of impulsive noise [19], whereby the distribution amplitude of the impulsive noise is modeled using Gaussian process and the probability of impulsive noise occurrence is modeled using Poisson process. In this regard the effective variance of Middleton Class A noise model is given as,

$$\sigma_I^2 = K\frac{\sigma_g^2}{A}$$

where the parameter A represents the density of the impulse noise and K represents the power of the impulse noise, σ_I^2 is the variance of the impulsive noise and σ_g^2 is the variance of the AWGN, which is represented as,

$$\sigma_g^2 = \frac{N_0}{2}$$

where N_0 represent the noise power spectral density, and the division by 2 indicates a double-sided power spectral density.

D. Signal analysis of PLC ACO-OFDM

To analyse the transmitted signal through the hybrid PLC/ACO-OFDM system, the communications system in Fig. 1 will be referred to. Information is transmitted on about quarter of the subcarriers in ACO-OFDM. The frequency domain signal to the input of the inverse fast Fourier transform (IFFT), D, contains only odd components such that, $D = [0, D_1, 0, D_3, \ldots, D_{N-1}]$, and again this signal is constrained to have Hermitian symmetry property given as,

$$D_m = D^*_{N-m}, \; for \;\; 0 < m < N/2, \qquad (1)$$

where D* represents the complex conjugate of D.

After performing the Hermitian symmetry operation on D, the frequency domain signal then undergoes nulling (or setting to zero) of even subcarriers, to produce signal X, according to the expression below

$$X(k) = \begin{cases} D^*(k) & , k \quad odd \\ 0 & , k \quad even \end{cases}$$

The signal at the output of the IFFT is real and this is guaranteed by the use of Hermitian symmetry property. In order to satisfy the Hermitian symmetry condition the component D_0 and $D_{N/2}$ in (1) are set to zero, $D_0 = D_{N/2} = 0$. The resulting time domain signal is real and given as,

$$s(n) = \frac{1}{N}\sum_{k=0}^{N-1} X(k) exp(j2\pi\frac{kn}{N}) \qquad (2)$$

for $n = 0, 1, \ldots, N-1$,

where N is the number of subcarriers and $X(k)$ is the k^{th} subcarrier of signal X. In practice, for large values of N the signal $s(n)$ is usually modelled as Gaussian random noise with zero mean and variance,

$$\delta^2 = E\{s^2\}$$

where $s(n)$ is the n^{th} time domain sample of signal s and $E\{\cdot\}$ is the expectation operation.

The time domain signal, $s(n)$, appearing at the output of the IFFT is then passed through a PLC channel whereby the impulsive noise, n_I and background noise modelled as AWGN, n_A are the limiting noise factor in terms of the performance of the channel. The resulting signal x from the

PLC channel is now contaminated with a mixture of impulsive noise and background noise given as,

$$x = s + n_I + n_A . \quad (3)$$

It is worth noting that this signal x is still bipolar and needs to be adjusted appropriately in this regard to make it unipolar. This process concludes the ACO-OFDM technique resulting in a time domain signal x_u denoted as,

$$x_u = \begin{cases} x(k) & , \ if \quad x(k) > 0 \\ 0 & , \ if \quad x(k) \leq 0 \end{cases} . \quad (4)$$

With consideration of Hermitian symmetry and the fact that ACO-OFDM only sends information only on the odd components then the data carrying subcarriers are $N_d = (N - 2)/4$.

E. Combating noise

To mitigate the impulsive noise contribution, the time domain signal will need to be pre-processed using memoryless nonlinearity techniques [20], [21]. In this paper we propose two most common techniques known to deal with impulsive noise, where by the first approach is to null (blank) the PLC OFDM signal and the second approach is clipping the signal. These techniques are designed based on the fact that impulsive noise amplitudes are usually higher than that of the actual PLC OFDM signal. In this regards a threshold T_c is used as a set point to distinguish between the samples affected by impulse noise and those not affected. Considering the signal, x_u, at (4) the nulling and clipping process is then denoted by (5) and (6) respectively,

$$x_{PLC-VLC}(k) = \begin{cases} x_u(k) & , x_u(k) \leq T_c \\ 0 & , x_u(k) > T_c \end{cases} \quad (5)$$

And,

$$x_{PLC-VLC}(k) = \begin{cases} x_u(k) & , x_u(k) \leq T_c \\ T_c & , x_u(k) > T_c \end{cases} \quad (6)$$

At this point we have managed to achieve a signal integrating both the PLC and VLC channel, $x_{PLC-VLC}(k)$, at the same time combating the impulsive noise using traditional techniques of nulling and or clipping respectively. These techniques are usually used to identify signal amplitudes that are larger than T_c or smaller than $-T_c$ and apply either nulling or clipping technique depending on the one used at that point in time. Therefore, this means both the negative and the positive samples higher than the given threshold, T_c are nulled or clipped. However, that is not the case in this proposed system, in (5) and (6) it should be noted that only the positive samples greater than Tc are nulled or clipped and this is because the ACO-OFDM automatically detect any negative samples regardless of whether they are affected by impulsive noise or not and deal with them accordingly as indicated in (4).

The signal $x_{PLC-VLC}$ again experience background noise (AWGN) due to VLC channel represented as,

$$y = x_{PLC-VLC} + n_A$$

III. RESULTS AND DISCUSSION

In this simulation the PLC channel takes into account two noise models, impulsive and AWGN, while the VLC channel only takes into account the AWGN as shown in figure 1. Before analysis of the results in detail, it is important to note this parameter which will be used to carry out the simulations, defined as follows: A is the probability of impulsive noise, K is the impulsive noise power, T_c is the threshold, σ_g^2 and σ_I^2 are the variances of AWGN and impulsive noise. Section C clearly indicate the relationship between the variances and the impulsive noise probability, A. In this regard, for all simulations' different values of T_c will be varied and tested for different scenarios of A and K in the PLC-VLC channel.

The simulation results presented, demonstrate the best threshold T_c under different conditions of impulsive noise for the PLC-VLC system presented in Fig. 1. Different values of T_c were experimented on for different impulsive noise scenarios such that the optimum T_c can be deduced. It is to be noted that applying the ACO-OFDM technique as demonstrated in (4), the negative samples above T_c will be identified and set to zero (nulling) and positive samples above T_c will be identified and set to zero according to (5).

The results can be discussed in groups as follows. In Fig. 2, the power of the impulsive noise is set to K = 10, while in Fig. 3 it is set to K = 100 and A = 0.1. In Fig. 2, at BER of 10^{-4}, the SNR for T_c = 2 is about 32 dB, while for T_c = 4 is about 29 dB. Looking at the similar scenario for Fig. 3 it can be seen that at BER of 10^{-4}, the SNR for T_c = 2 is about 41 dB and for T_c = 4 is about 39 dB. It is clear from the simulations results that Fig. 3 yields best performance at low SNR when compared with Fig .2. While on the high SNR Fig. 2 yields best performance when compared to Fig .3. However, in both cases Fig. 2 and Fig. 3 presents common property where by the optimum Tc for low SNR is 2 and for high SNR is 4.

In the next set of simulations represented in Fig. 4 and Fig. 5, the parameters are set similar to Fig. 2 and Fig. 3, except that this time the impulsive noise probability is A = 0,01 and the impulsive noise power is set to K = 10. In Fig. 4, at BER of 10^{-4}, the SNR for T_c = 2 is about 15 dB, while for T_c = 4 is about 14 dB. Again in Fig. 5 it can be seen that at BER of 10^{-4}, the SNR for T_c = 2 is about 15 dB and for T_c = 4 is about 14 dB. This shows that applying a threshold of T_c = 4 results in SNR which is 1 dB better than when a threshold T_c = 2. The results obtained in Fig. 4 and Fig. 5 show that in both cases the best T_c is achieved at T_c = 4, the drastic improvement in the SNR for Fig. 4 and Fig. 5 as opposed to Fig. 2 and Fig. 3 is due to reduced probability of noise occurrence, A. Furthermore Fig. 4 and Fig. 5 clearly indicate that increasing the impulsive noise power, K does not necessarily influence the performance of the system. The reason for the low dependency on K, is that once the impulsive noise amplitudes are detected using threshold T_c, they are nulled, rendering the power of the impulsive noise ineffective. It is to be noted that the number of nulling operations depends on the parameter A, this means an increase in A will result in an increase in the number of nulls in the signal. Hence the best system performance is achieved when the value of A is low as in Fig. 4 and Fig. 5.

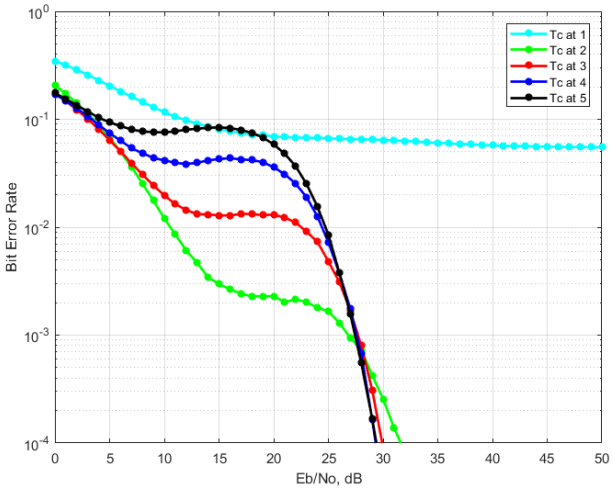


Fig. 2: Performance of Hybrid PLC-VLC based on ACO-OFDM (BPSK): A = 0.1, K = 10 and T_c {1, 2, 3, 4, 5}

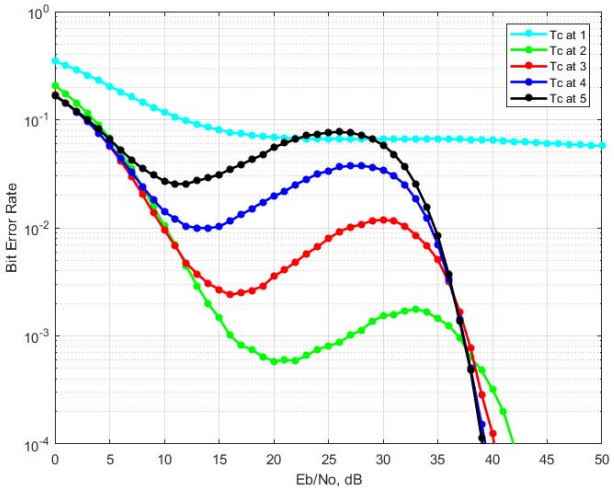


Fig. 3: Performance of Hybrid PLC-VLC based on ACO-OFDM (BPSK): A = 0.1, K = 100 and *Tc* {1, 2, 3, 4, 5}

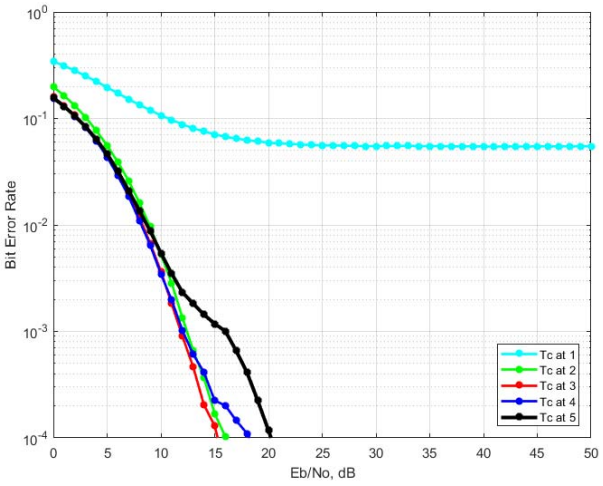


Fig. 4: Performance of Hybrid PLC-VLC based on ACO-OFDM (BPSK): A = 0.01, K = 10 and T_c {1, 2, 3, 4, 5}

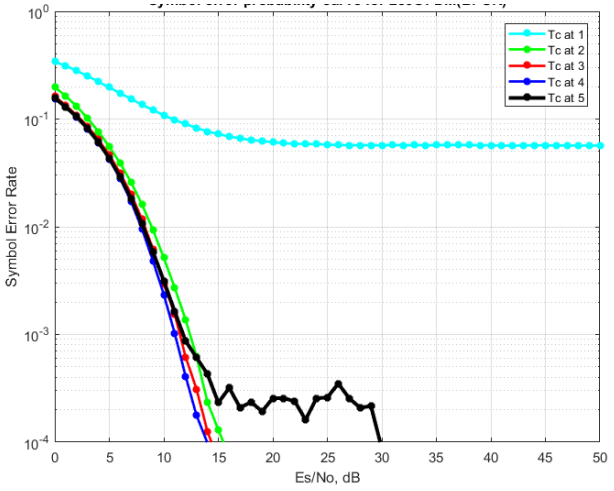


Fig. 5: Performance of Hybrid PLC-VLC based on ACO-OFDM (BPSK): A = 0.01, K = 100 and T_c {1, 2, 3, 4, 5}

IV. Conclusion

In this paper simulation results of impulsive noise evaluation to be used in PLC-VLC communication system based on ACO-OFDM is presented. It is found that the overall performance of this communication system is mainly influenced by the impulsive noise parameters A and K. It is shown from the BER that the system performance is highly dependent on the impulsive noise probability, A and less dependent on the impulsive noise power, K. The large probability of A results in an increased number of null operations to be performed on the OFDM signal. The low dependency on the noise power K is due to the fact that once any impulsive noise amplitudes are detected based on the set threshold, T_c, they are nulled hence rendering the impulsive noise power ineffective. The practical implementation of this system can be done for further research and to validate the system performance in real life. Simulations using the Hartley transform can be further tested against the results presented in this paper, replacing the need for Hermitian symmetry. In addition, it will be interesting to compare the results in this work with those investigated in [22], where an integration of a PLC system with VLC employing DCO-OFDM was presented.

References

[1] S. Dimitrov and H. Haas, "Information rate of OFDM-based optical wireless communication systems with nonlinear distortion," J. of Lightwave Tech.,31, (6),918-929,2013.

[2] M. S. Mossaad, S. Hranilovic, and L. Lampe, "Amplify-and-forward integration of power line and visible light communications," in IEEE Global Conf. on Signal and Inf. Process. (GlobalSIP), Orlando, FL, USA, 14-16 Dec. 2015, pp. 1322–1326.

[3] A.R. Ndjiongue, H.C. Ferreira, and T. Shongwe, "Inter-building PLC-VLC integration based on PSK and CSK techniques," in 20th IEEE ISPLC Conf., Bottrop, Germany, 20-23 Mar. 2016, pp. 31–36.

[4] "IEEE standard for local and metropolitan area networks–part 15.7: Short-range wireless optical communication using visible light," IEEE Std 802.15.7-2011, pp. 1 –309, 6 2011.

[5] S. Rajagopal, R. Roberts, and S. Lim, "IEEE 802.15.7 visible light communication: modulation schemes and dimming support," IEEE Commun. Mag., vol. 50, no. 3, pp. 72–82, 2012.

[6] A. R. Ndjiongue, T. Shongwe, H. C. Ferreira, T. N. Ngatched, and A. H. Vinck, "Cascaded PLC-VLC channel using OFDM and CSK techniques," in IEEE GLOBECOM Conf., San Diego, CA, USA,6-10 Dec. 2015, pp. 1–6.

[7] H. Ferreira, L. Lampe, J. Newbury, and T. Swart (Editors), Power Line Communications: Theory and Applications for Narrowband and Broadband Communications over Power Lines. John Wiley & Sons, 2010.

[8] ITU-T Rec. G.9960 and G.9961, "Unified high-speed wire-line based home networking transceivers," 2010.

[9] S. Galli and O. Logvinov, "Recent developments in the standardization of power line communications within the IEEE," IEEE Commun. Mag., vol. 46, no. 7, pp. 64–71, Jul. 2008.

[10] V. Oksman and S. Galli, "G.hn: the new ITU-T home networking standard," IEEE Commun. Mag., vol. 47, no. 10, pp. 138–145, Oct. 2009.

[11] IEEE Std 1901-2010, "IEEE standard for broadband over power line networks: Medium access control and physical layer specifications,"2010.

[12] A. R. Ndjiongue, T. Shongwe, and H. C. Ferreira, "Closed-Form BER Expression for HSV-Based MPSK-CSK systems," IEEE Commun. Lett., vol. 21, no. 5, pp. 1023–1026, 2017.

[13] A. R. Ndjiongue, T. Shongwe, H. C. Ferreira, T. N. Ngatched, and A. H. Vinck, "PSK to CSK mapping for hybrid systems involving the radio frequency and the visible spectrum," Telecommun. Sys., vol. 64, no. 1, pp. 173–192, 2017.

[14] H. C. Ferreira, H. M. Grove, O. Hooijen, and A. H. Vinck , Power line communications. Wiley Online Library, 2010.

[15] M. Korki, N. Hosseinzadeh, H. L. Vu, T. Moazzeni, and C. H. Foh, "A channel model for power line communication in the smart grid," in IEEE/PES Power Sys. Conf. and Exposition (PSCE), Phoenix, AZ, USA, 20-23 Mar. 2011, pp. 1–7.

[16] E. Monteiro and S. Hranilovic, "Constellation design for colorshift keying using interior point methods," in IEEE GLOBECOM Workshops (GC Wkshps), Anaheim, CA, USA, 3-7 Dec. 2012, pp. 1224–1228.

[17] R. J. Drost and B. M. Sadler, "Constellation design for channel precompensation in multi-wavelength visible light communications," IEEE Trans. Commun., vol. 62, no. 6, pp. 1995–2005, 2014.

[18] J. A. Bingham, "Multicarrier modulation for data transmission: An idea whose time has come," IEEE Commun. Mag., vol. 28, no. 5, pp. 5–14, 1990.

[19] T. Shongwe, A. H. Vinck, and H. C. Ferreira, "On impulse noise and its models," in 18th IEEE ISPLC Conf., Glasgow, UK, 30 Mar.-2 Apr. 2014, pp. 12–17.

[20] S. Zhidkov, "Analysis and comparison of several simple impulsive noise mitigation schemes for OFDM receivers", IEEE Trans. Commun., Vol. 56, No. 1, pp. 5-9, Jan. 2008.

[21] C. R. Nokes, O. P. Haffenden, J. V. Mitchell, A. P. Robinson, J. H. Stott and A. Wiewiorka, "Detection and removal of clipping in multicarrier receivers", European patent application EP1043874, filed Jul. 4 2000, issued Oct. 11 2000.

[22] M.D. Kubjana, T. Shongwe and A.R Ndjiongue, "Impulsive Noise Evaluation on PLC-VLC based on DCO-OFDM integration based on," in 11th IEEE/IET CSNDSP Conf., Budapest, Hungary, 18-20 Jul. 2018, pp. 1–6.

System for Monitoring and Control of the Baxter Robot

Yasen Yordanov
Faculty of Automatics
Technical University of Sofia, Bulgaria
valerim@tu-sofia.bg

Ognyan Nakov
Faculty of Computer Systems and Technologies
Technical University of Sofia, Bulgaria
nakov@tu-sofia.bg

Valeri Mladenov
Faculty of Automatics
Technical University of Sofia, Bulgaria
valerim@tu-sofia.bg

***Abstract*—This scientific report details how a software and hardware system can be created to allow the surveillance, access and control of a Baxter industrial robot from any location connected to the Internet. Baxter is being aimed at the nearly 270,000 small to midsize manufacturers, which have five hundred or fewer employees - companies of that size are unlikely to be able to invest hundreds of thousands of dollars into robots that require a redesign of their workspace and IT personnel to run them. Commonly used, the robot can handle many different industry tasks – co-packing and end-of-line packaging, pick and place operations, plastic injection, printed circuit board handling and many others. It's important to noted, that the platform on which the control software works is Microsoft Windows. The programming language used to write the code for the system is Python 2.7. For monitoring a robot when executing a command, a high-quality camera - AUSDOM AW615 1080p - is also used in the system.**

Keywords— Baxter, Python, Software, WebCam

I. INTRODUCTION

The Baxter robot has many and different applications [1]. Thanks to its embedded cameras, a team from New Zealand has managed to develop a project in which the robot plays autonomously chess against humans [2]. This is achieved with the camera, positioned in one of the robot's arms, that perceive the game state, an open-source chess engine [3] that computes the next move, and a mechatronics subsystem with a 7-DOF arm [4] that handles the pieces. Also there are many different models of the robot, that are developed to ease the it's use. The dynamic model for example is important, because he helps to find the relationship between the joint actuator torques and the resulting motion in the robot. The models is developed by a team from the United Kingdom and uses Lagrange Euler (L-E) equations [5]. In addition, the team has also developed and presented the kinematic model of the robot.

In general, the Baxter robot is required to have a separate computer on which the Robotic Operation System (ROS) [6] is installed. It accesses the robot and its main functions. The problem is that it is Linux [7] based, so Windows[8] computers can not control the robot. Another problem is that the computer and the robot must be connected in a common network so that they can send and receive commands to each other. Of course, difficulties in working with the robot appear because of its size - 182 cm height and weight 138 kg. with the tray. With these dimensions, dragging the robot for demonstrations and training becomes a difficult task. The software developed & described in the report solves these problems. It is written for MS Windows operating systems. A socket is used to connect to the computer running the robot. The only other necessary hardware is a camera through which the operator monitors the proper execution of the commands provided by the software. Thanks to the code developed, the robot can perform different movements with its arms - stretched up, stretched to the side and then back to neutral position, and can also move its head - left, right and return to neutral position. Below, the report describes both the software packages involved in writing the system code and the hardware used. There is also a special insight into testing the system - in the chapter Tests and Validation.

II. SOFTWARE

2.1. Introduction

The Baxter robot is only compatible with Python [9] version 2.7 - due to restrictions associated with the use of "ROS". ROS, in addition to managing the robot, is only compatible with Linux based operating systems. In order to manage the robot through Windows, we use the python WebSocket library. In order to communicate between the two operating systems, the command is first sent through the websocket from the user interface developed by the Tkinter library, and then processed by the Ubuntu machine, which, depending on the information, sends the required execution command to the robot. The Windows machine in the system plays a role as a client, and the Linux one on a server. Before the software is used, it is necessary to create a working environment. This is done by using the command catkin make. The necessary packages can then be used. This command is also used to configure the path to the code so that it can be imported. In case other ROS packages are used, the paths to the libraries that will be used must be dynamically attached. Different classes are used to wrap the envelopes.

2.2 Libraries used in the system

2.2.1 Python "tkinter"

This is the library that was used to create the user's graphical interface, which is deployed on the Windows personal computer. Some of the key functions used to write the code are: button_connect (), button_config () and button.grid ().

2.2.2. Python "sys"

The library is used to operate the system with the interpreter. For example, to get out of a function with exit (), we need this library.

2.2.3. Python "time"

Provides various time-related features. For the proper operation of the library, it's important to get to know the operating system (OS) well, because each OS has a different implementation of the time-processing functions. The function, used in the system, to create delay is time.sleep().

2.2.4. ROS "Baxter_interface"

This is a library that initializes the particular part of the Baxter robot that will be used. It is possible to initiate more than one part of the robot, which allows for more complex operations.

2.2.5. ROS "Rospy"

This is the library that makes it possible to interact with the properties of ROS faster and easier.

2.2.6. WebSocket Library

WebSocket is a library that we use to create a communication protocol for a persistent, bi-directional, full duplex TCP connection from a client to a server. A WebSocket connection is initiated by sending a WebSocket handshake request from a HTTP connection to a server to upgrade the connection. Along with the upgrade request header, the handshake request includes a 64-bit Sec-WebSocket-Key header. The server responds with a hash of the key in a Sec-Websocket-Auth header. This header exchange prevents caching proxy from resending previous WebSocket exchanges. From that point, the connection is binary and does not conform to HTTP protocol. The WebSocket API is an advanced technology that makes it possible to open a two-way interactive communication session between a client and a server.

2.3 Robot Operational System (ROS)

Robot Operational System (ROS) is used to communicate with the robot. Once a command is sent from the user interface through the WebSocket, this information is processed, and with ROS the necessary connection with the robot's hardware is created. Then the commands set by the software are executed. ROS works with python version 2.7, scripts can be written interactively via a terminal or in separate files - as in the case of system development. ROS is sometimes called a "meta operating system" because it performs many of the functions of an operating system. One of its main purpose is to provide communication between the user, the Ubuntu OS and, of course, Baxter.

As with any operating system, the benefit of ROS is the hardware abstraction and low-level control of Baxter without the Baxter user knowing all the details of the robot. The robot, as well as the control station, must be connected to the same internet network so they can communicate with each other. ROS has several visualization and easy-to-read controls. For example, MoveIt is a planning framework for movements and operations. It makes it easier to take some of the necessary coordinates needed for proper positioning of the robot's hands and head.

III. HARDWARE

3.1 Introduction

Two types of hardware are used in the system - “Baxter robot” and a high-quality camera AUSDOM AW615 1080p. The camera is used used to track the robot state by an operator.

3.2 Baxter robot [3]

3.2.1 Physical specifications

The robot has 3’ 1” height without pedestal and between 5’ 10” and 6 ’3” with the adjustable pedestal. It’s maximum reach is1210 mm. The torso mounting plate diameter is 13.3” and it is used for mounting on table. It has body weight 165 lbs. without pedestal and 306 lbs. with pedestal. The robot has 14 Degrees of Freedom (DOF) - 7 per arm.

The pedestal footprint is with dimentions 36” × 32”. Baxter’s max Payload is 5 lb / 2.2 kg and he has gripping torque maximum 10 lb /4.4 kg.

3.2.2 Computer and sensor specifications

The robot has 3rd Gen Intel Core i7-3770 (8MB, 3.4GHz) processor [10] w/HD4000 Graphics, 4GB memory (NON-ECC, 1600MHZ DDR), 128GB Solid State Drive Storage. Baxter’s camera has max resolution 1280 x 800 pixels, it is effective resolution is 640 x 400 pixels and has 30 frames per second frame rate. The camera’s focal length is 1.2 mm. The robot has screen on his head with screen resolution 1024 x 600 pixels. On his head he has also infrared sensors [11] with range between 1.5 – 15 in / 4 – 40 cm.

3.2.3 Electrical specifications

The supply voltage of the robot is 120 volts alternating current with rated current - 6 amps. The robot can operate on battery – he has internal DC-to-120V AC Inverter [12]. It is important to note that the Baxter robot has an internal PC, which cannot be powered directly off of 24V DC. Baxter has standard 120VAC power interface. Robot power bus and internal PC both have “universal” power supplies and support 90 – 264V AC (47 – 63Hz). It is maximum consumption is 6A at 120V AC, 720W max per unit, the maximum efficiency varry between 87% to 92%.

For power supply the robot uses medical-grade DC switching power supply for robot power bus. Baxter has also tolerance to Sags – they are tolerated to 90V but sustained interruption will require manual power-up. The Voltage Flicker holdup time is 20ms and the voltage unbalance supports single phase operation only.

3.3 Camera AUSDOM AW615 1080p

3.3.1 Details

The camera has 123g. weight and dimensions 66*35*135mm

3.2.2. Lens Spec

They have focus from 30cm to infinity. The material, used for creating them is PC-6100D1, 4-layer film-coated glass lens group. They have 65° horizontal viewing angle and support zooming.

3.2.3. Microphone Spec

The microphone is built-in. It has S.P.L 32dB±2Db and it is omni-directional

3.2.4. Hardware Spec

The camera's chipset is P269+MA1080, the control IC is P269 (REALTEK RTS5822) QFN46 and the sensor is MA1080(5B3) CSP-48 1/4.5 CMOS "SAMSUNG". The user can choose between USB3.0 [13]/USB2.0/1.1 for connection interface. The camera's power consumption is ≤220MA and it supports different Operating Systems: Windows 7, Windows XP2, Windows 8 and Windows Vista

3.2.5. Video/Image Spec

The video resolution is maximum 1920*1080 5fps and minimum is 160*120 30fps [14]. There are 2 video modes: the default one is YUY2 with a maximum resolution of 1920*1080, 30fps and MJPG format and a minimum resolution of 160*120, 30fps. The photo resolution is 1920*1080 and the photo format is JPG [15].

IV. Experiments and Results

4.1 Robot tests

The system allows the user to execute up six different commands - three for movement of the head and three for movement of the robot's arms:

- Rotate the robot's head to the left
- Rotate the head of the robot to the right
- Neutral position of the robot's head
- Hands stretched upwards
- Hands stretched sideways
- Neutral hand position

The test set-up is show on Figure 1.

Fig. 1. View of the system's test set-up

The figures below show the movements of the hands of Baxter and the movement of the robot's head.

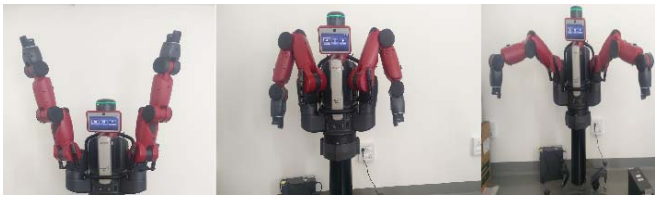


Fig. 2. Examples of execution of a command for arm movement set by the GUI

Fig. 3. Examples of execution of a command for head movement set by the GUI

4.2. Software testing

The developed management software has seven buttons - one for each movement performed by the robot, and one for the connection between the individual workstations. The software also displays the information received from the camera.

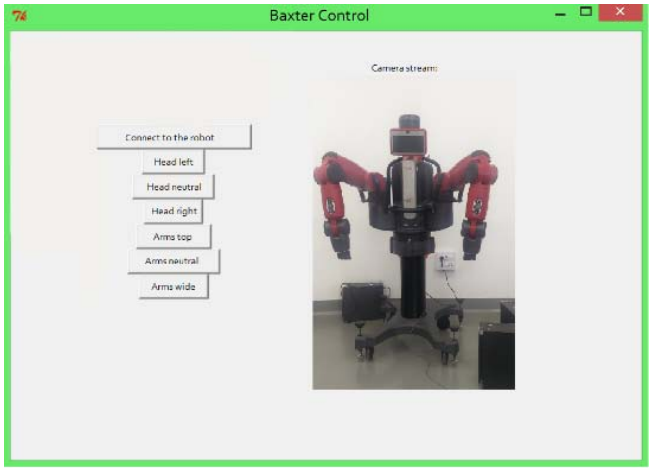


Fig. 4. View of robot's control software

4.3 Testing the performance of the system

The system has been tested for its performance. Table 1 below describes:

- Command - the command sent from the user interface to the robot for execution
- Reaction time - the time between the initialization of the command and its completion - measured in seconds
- Status - a column for whether the command has been executed

TABLE 1 - REACTION TIME OF THE SYSTEM

Command	Reaction time (s)	Status
Head position neutral	1.35	OK
Head position neutral 2	1.19	OK
Head position left	1.32	OK
Head position left 2	1.41	OK
Head position right	1.21	OK
Head position right	1.23	OK
Arm position neutral	3.21	OK
Arm position neutral 2	3.32	OK
Arm position straight up	2.53	OK
Arm position straight up 2	2.46	OK
Arm position wide	2.37	OK
Arm position wide 2	2.39	OK

V. CONCLUSION

As a result from the tests, it is clear that the system performs all of the required tasks - running on Windows OS, connecting to the Linux workstation, controlling the robot's actions and displaying real-time performance. Also the overall reaction time of the system is very fast. Still the reaction time depends more or less on the internet connection of the user, so the results received by different users may differ from the experimental data given in this paper. The future development of the system includes improvements to the graphical interface - modernizing the view as well as adding more execution commands to be performed by the robot. Furthermore, the camera will be replaced with Microsoft Kinect in order for the robot to be controlled by the actions of a person standing in front of the robot's sensor by following the person's movements.

ACKNOWLEDGMENT

This work is supported with project "Development of mobile android application for control of industrial robot Baxter" № 182ПД0015-08 / session 2018 of the Research and Development Sector of the Technical University of Sofia.

REFERENCES

[1] S. Cremer, L. Mastromoro and D. O. Popa, "On the performance of the Baxter research robot," 2016 IEEE International Symposium on Assembly and Manufacturing (ISAM), Fort Worth, TX, 2016, pp. 106-111.

[2] A. T. Chen and K. I. Wang, "Computer vision based chess playing capabilities for the Baxter humanoid robot," 2016 2nd International Conference on Control, Automation and Robotics (ICCAR), Hong Kong, 2016, pp. 11-14.

[3] M. Levene and J. Bar-Ilan, "Comparing Typical Opening Move Choices Made by Humans and Chess Engines," in The Computer Journal, vol. 50, no. 5, pp. 567-573, Sept. 2007.

[4] T. Zhao, J. Yuan, M. Zhao and D. Tan, "Research on the Kinematics and Dynamics of a 7-DOF Arm of Humanoid Robot," 2006 IEEE International Conference on Robotics and Biomimetics, Kunming, 2006, pp. 1553-1558.

[5] A. Smith, C. Yang, C. Li, H. Ma and L. Zhao, "Development of a dynamics model for the Baxter robot," 2016 IEEE International Conference on Mechatronics and Automation, Harbin, 2016, pp. 1244-1249.

[6] H. Wei, Z. Huang, Q. Yu, M. Liu, Y. Guan and J. Tan, "RGMP-ROS: A real-time ROS architecture of hybrid RTOS and GPOS on multi-core processor," 2014 IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, 2014, pp. 2482-2487.

[7] M. Kong, J. Li and Wang Fengming, "Study on educational mode of Linux majors in colleges," 2010 International Conference on Artificial Intelligence and Education (ICAIE), Hangzhou, 2010, pp. 623-626.

[8] H. Upadhyay, H. A. Gohel, A. Pons and L. Lagos, "Windows Virtualization Architecture For Cyber Threats Detection," 2018 1st International Conference on Data Intelligence and Security (ICDIS), South Padre Island, TX, 2018, pp. 119-122.

[9] B. A. Malloy and J. F. Power, "Quantifying the Transition from Python 2 to 3: An Empirical Study of Python Applications," 2017 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM), Toronto, ON, 2017, pp. 314-323.

[10] L. Dake, C. Zhaoyun and W. Wei, "Trends of communication processors," in China Communications, vol. 13, no. 1, pp. 1-16, Jan. 2016.

[11] K. Kudlaty, A. Purde and A. W. Koch, "Development of an infrared sensor for on-line analysis of lubricant deterioration," SENSORS, 2003 IEEE, Toronto, Ont., 2003, pp. 903-908 Vol.2.

[12] K. Rahimi, A. N. Motlagh and M. Pakdel, "A new soft-switched ZCZVT DC-AC inverter," 2009 IEEE Vehicle Power and Propulsion Conference, Dearborn, MI, 2009, pp. 1338-1344.

[13] P. Wüstner et al., "The use of USB 3.0 for fast data transfer in a PET detector," 2014 19th IEEE-NPSS Real Time Conference, Nara, 2014, pp. 1-2.

[14] M. Sultanoff, "A 100,000,000 Frame per Second Camera," in Journal of the Society of Motion Picture and Television Engineers, vol. 55, no. 2, pp. 158-166, Aug. 1950.

[15] P. B. Pawar, P. K. Kawadkar, M. Nagle and P. K. Ambare, "Analysis of signature and signature free bufferoverflow detection for gif and jpg format," 2013 Tenth International Conference on Wireless and Optical Communications Networks (WOCN), Bhopal, 2013, pp. 1-5.

A Controllable Deflection Routing and Wavelength Assignment Algorithm in OBS Networks

Philani Khumalo
Faculty of Engineering,
Durban University of Technology
Durban, South Africa
khumalopk@elec.durban.gov.za

Bakhe Nleya
Faculty of Engineering
Durban University of Technology
Durban, South Africa
bmnleya@ieee.org

***Abstract*—Heterogeneous IoT enabled networks generally accommodate both jitter tolerant and intolerant traffic. Optical Burst Switched (OBS) backbone networks are handling the resultant volumes of such traffic by transmitting it in huge size chunks called bursts. Because of the lack or limited buffering capabilities within the core network, contentions as well as congestion may frequently occur and thus affecting overall supportable quality of service (QoS). Both contention and congestion will be characterized by frequent burst losses especially when traffic levels surge. The congestion is normally resolved by way of deflecting contending bursts to other less congested paths even though this may lead to differential delays incurred by bursts as they traverse the network. This will contribute to undesirable jitter that may ultimately compromise overall QoS. Noting that jitter is mostly caused by deflection routing which itself is a result of poor wavelength and routing assigning, in this paper we propose a controllable deflection routing (CDR) scheme that allows the deflection of bursts to alternate paths only after controller buffer preset thresholds are surpassed. In this way bursts intended for a common destination are always most likely to be routed on the same or least cost path end-to-end. We describe the scheme as well as compare its performance to other existing approaches. Both analytical and simulation results overall show that the proposed scheme does lower both congestion as well as jitter, thus also improving throughput as well as avoiding congestion on deflection paths.**

Keywords—Optical Burst Switching, jitter, deflection routing, congestion

I. INTRODUCTION

In the OBS domain primary concerns are in combating congestion as well as contention as bursts traverse the core network. In any given network various types of congestion e.g., nodal, CPU, path, may occur. Nodal congestion occurs when incidence traffic overwhelms the serving node. CPU congestion is as a result of too many computations that jam the main CPU scheduler. Path or link congestion is caused by excessive traffic attempting to traverse the same path. In the context of OBS networks, congestion thus can be caused by several factors such as contention; uneven distribution of traffic leading to localized traffic overload, as well as improper provisioning of available resources such as is in the case of routing and wavelength assignment (RWA). The presence of buffering capabilities at edge nodes makes it easy to combat edge congestion. Path congestion can be alleviated by way of dimensioning the available network resources such as wavelengths and links such that traffic is uniformly distributed throughout the network [1].

It is noted that contention will always occur at interior nodes when more than one data bursts utilizing the same wavelength overlap in time at the same single output port. Because of the bufferless nature of such networks in their interior, different approaches are adopted such to alleviate as well as combat contention. Primarily the contention resolution mechanisms can be implemented at space, wavelength, or time domains. At wavelength domain level, wavelength converters (WCs) may be used occasionally to resolve the contention by translating one of the contending wavelengths to a different value. In so doing the network's performance improves. In the time domain contention resolution is implanted/effected by introducing fiber delay lines (FDLs), to temporarily delay one or more of the contending bursts until such time that the output port becomes available. In space domain, deflection routing is introduced to resolve any contention occurrences in which one of the contending bursts can be deflected to an alternate port as well as route. In this way both congestion as well as contention are distributed to other routes rather than being concentrated on a single one and in the process the network's general performance improves. Nevertheless, it should be noted that deflection routing also has several draw backs, notably that it can accelerate contention as well as congestion on the deflection paths. Its performance is largely influenced by the general network topology it may not feature effectively where the number of candidate deflection paths are relatively small. Furthermore, it can also contribute to differential delays or jitter for successive bursts destined for the same receiver as the deflected bursts might either take a longer a shorter path than their non-deflected counterparts., It is thus imperative that the deflection routing itself must be controlled [3]-[5].

It is on the strength of the earlier cited weakness that in this paper, we propose a controllable deflection routing scheme which couples with a simple wavelength and routing assignment (WRA) algorithm to enhance network minimizing both contention and congestion. The scheme attempts as much as possible to route the contending bursts along paths that have been chosen based on the minimization of performance measures such as delay and blocking. The scheme aims at controlling deflection traffic by way of selective path routing upon congestion onset. It is backed by a very simplified distributed RWA approach that ensures minimal contention in the primary (original) chosen route(s). Notably a distinct feature of the proposed scheme is that it allows the deflected bursts to traverse further via deflection routes optimized for improved performance in terms of delay and blocking. The candidate deflection routes are themselves dynamically classified according to QoS constraints (e.g., blocking and delay) they can support [6][7].

The paper is outlined as follows: In the next section we elaborate on the proposed the algorithm in more detail. In section three we model the controllable deflection routing queuing model. Finally, section 4 presents both analytical as well as simulation results pertaining to proposed scheme. Conclusions are drawn in section 5.

II. PROPOSED CONTROLLABLE DEFLECTIONC ROUTING (CDR) SCHEME

We commence the section by describing the proposed scheme. Figure 1 depicts a generalized architecture of an OBS switch which comprises several input and output wavelength division multiplexed (WDM) link ports.

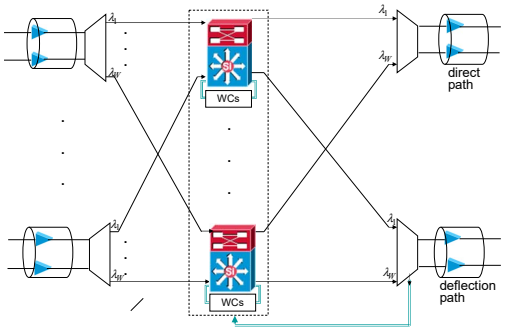


Figure 1. Switch architecture with WCs

In general, wavelength light paths from input fibers are demultiplexed prior to switching to the desired output ports. In the event of contention, one of the contending data bursts is deflected to an alternate route. Periodic global re-optimizing of candidate deflection routes based on most recently exchanged contention as well as congestion status updates from other nodes is necessary. In the event that the network management system reports contention as well as wavelength congestion on the deflected route or their imminence, the contending burst may be converted to any other available wavelength by an WC.

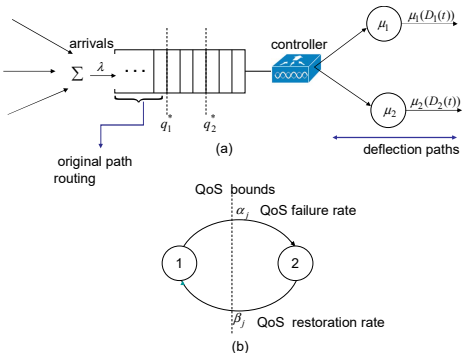
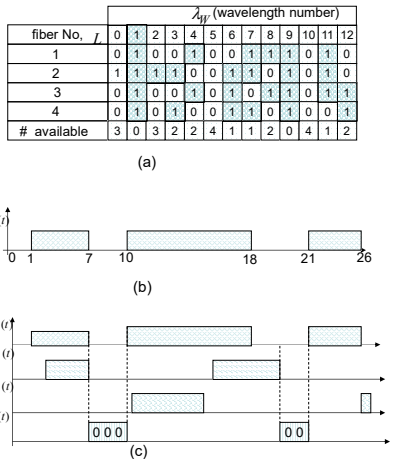


Figure 2. Queuing model

This updating interval is carefully selected in accordance with the computing power capabilities of the node so as not to cause nodal computational congestion. As can be seen in Figure 1, the switch fabric can only accommodate a limited number of both optical links as well as wavelengths.



fiber No, l	λ_W (wavelength number)											
	0	1	2	3	4	5	6	7	8	9	10	11	12
1	0	1	0	0	1	0	0	1	1	1	0	1	0
2	1	1	1	1	0	0	1	1	0	1	0	1	0
3	0	1	0	0	1	0	1	0	1	1	0	1	1
4	0	1	0	1	0	0	1	1	0	1	0	0	1
# available	3	0	3	2	2	4	1	1	2	0	4	1	2

Figure 3. Wavelength management; (a). Example link state data structure , (b) example single wavelength occupancy state sequence, (c). Example concatenated wavelength occupancy state.

The number of input/output switch pairs tally with the number of shared WCs. A key feature of this switch architecture is that the choice as well as usage of deflection paths is controlled and by all means it will always thrive to route bursts intended for a common receiver/destination pair on the same path. Only in the event that routing on the same path is not possible, is the contending burst converted to another wavelength or, worst case it is deflected on to a least cost route in terms of delay and blocking QoS metrics. The scenario just described is further illustrated by the queuing model of Figure 2. All arriving bursts are served on a FCFS service discipline policy. Path server #1 queue represents a deflection path that offers minimal QoS degradation in terms of blocking and delay. Contending burst will be dispatched to server 1 queue only if the buffer threshold has exceeded a threshold state q_1.Path server #2 represents an alternate path which is utilized only when the controller buffer's threshold has exceeded q_2. Otherwise the original path is always chosen. Neither path can be expected to consistently meet its QoS expectations hence in general, we define, α as a given path's rate of exiting its QoS bounds, and similarly β would be the rate at restoring it to within bounds. This transition state is shown in Figure 2. (b).

In addition, also key to alleviating both contention and wavelength congestion is affective RWA. We propose a simplified RWA method which evenly distributes the number of available wavelengths on all fibers as well as links. A network routing map (NRM) together with simplex signaling are assumed. Each node furnishes as well as advertises the following static information to the NRM:

- candidate routes as well as overall network resources state to all destinations as example illustrated in Figure 3 (a).
- sum of available links as well as fibers(wavelengths).
- Each node also provides end-to end link occupancy states for all possible links from it to all other destinations.

An example individual fiber wavelength occupancy at each node is also illustrated in Figure 3(b). All this information is

dynamic hence it has to be updated periodically at an interval ΔT_{update} on the NRM.

The concatenated wavelength occupancy state can be represented by O such that:

$$O = (t, st) \quad (1)$$

where t, is the start time and st is the state of the slot.

A single wavelength's occupancy state can be represented by a sequence vector of slots as follows:

$$O_\lambda(t) = \left[O_1, O_2, ..., O_n\right] \quad (2)$$

The state occupancy of a concatenated links (candidate light path) can be defined as:

$$O_L(t) = \left[O_{\lambda 1}(t) \oplus O_{\lambda 2}(t) \oplus, ..., \oplus O_{\lambda W}(t)\right] \quad (3)$$

where the operation $\oplus$ denotes a search algorithm for free wavelengths along the links.

We can formulate the key deflection routing problem primarily as a function of the node configuration, general network topology as well as a set of QoS related attributes such as node and link resources [8]-[15].

If the physical network is denoted as $G(N, L)$, where N is the number of nodes comprising it and L is the set of links interconnecting the nodes.

Each link $L_{i,j}$ has a total of W_{ij} wavelengths each with capacity C.

Each network node n, $(n = \overline{1, N})$ has $P_n^{in}(t)$ and $P_n^{out}(t)$ ports. We define a source (s) and a destination (d) pair as well as an associated burst arrival rate $\lambda_{i,j}^{sd} \in \Lambda$,at the switch queue. We also define $\lambda_{s_k} d_k$ to represent the average flow of bursts belonging to class k type traffic. We thus can define:

$$x_{ij} = \begin{cases} 1, \text{ if deflection route includes, link } L_{i,j} \\ 0, \text{ otherwise} \end{cases} \quad i, j = \overline{1, N},\ i \neq j \quad (1)$$

Since one light path can be set up at each node we thus have:

$$\sum_{\Lambda, j \in N} \mathrm{x_{ij}} \leq P_i^{out}(t), \quad \sum_{\Lambda, i \in N} \mathrm{x_{ij}} \leq P_j^{in}(t) \quad (2)$$

Thus, the traffic demand $\lambda_{s_k} d_k$ deflected from node i to j is:

$$\lambda_{i,j}^{s_k, d_k} \in \left\{0, \lambda_{s_k d_k}\right\} \ \forall\ \mathrm{i,j} \in N \quad (3)$$

The aggregated one-way flow from node i to j associated with the k traffic demand is:

$$\lambda_{ij} = \sum_{s,d} \lambda_{ij}^{sd} + \lambda_{s_k d_k} \quad \forall_{i,j} \in N \quad (4)$$

Traffic from node i to j may not exceed the maximum capacity C hence we have:

$$\lambda_{ij} \leq W_{i,j} C \quad \forall_{i,j} \in N \quad (5)$$

If the same link $L_{i,j}$ is not associated with the k -the traffic type flow, then the previous equation becomes:

$$\lambda_{ij}^{s_k d_k} \leq x_{ij} \lambda_{s_k d_k} \quad \forall_{i,j} \in N \quad (6)$$

Finally, at each node the flow conservation constraint becomes:

$$\sum_i x_{ij} - \sum_j x_{ji} = \begin{cases} 1, & i = s_k \\ -1, & i = d_k \\ 0, & \text{otherwise} \end{cases} \quad \forall_{\mathrm{s_k}, d_k, i} \in N \quad (7)$$

Finally, if we let $D = \left\{D_{ij}\right\}$ represent the distance matrix as well as delay between nodes i and j we can thus summarize our key objective function as follows:

$$\text{Min } \gamma_{\mathrm{d}} \sum_{ij} x_{ij} D_{ij} + \gamma_b \left[\log\left[1 - \prod_{i,j}\left(1 - x_{ij} b_{ij}\right)\right]\right] \quad (8)$$

Where, γ_d and γ_b are the delay and blocking weights respectively. Collectively they are designated as a deflection path link cost factor.

$$c = f(\gamma_d, \gamma_b) \quad (9)$$

The key steps of the proposed CDR algorithm are summarized as follows:

i. Ingress (source) node dispatches a burst control packet (BCP) requesting an end-to-end connection to a specified egress (destination) node.
ii. the intermediate node processes the BCP together with those from other sources. If resources are available of the primary route (and is contention free), the burst will be accepted.
iii. However if contention is detected i.e. simultaneous requests for the same output ports and wavelengths, by two or more BCPs then the contention is resolved before actual burst arrival in one of the following ways:-

a. If the node is the sender, its BCP is discarded and retransmission at a later time is ordered.
b. The remaining bursts can either be assigned to the primary route, deflected to an alternate path, or worst case be discarded. This is done according to the set of rules in step iii:
iv. *assigned to the original path*: There exists two or more contending bursts all on transit. The node's controller is in state $q \prec q_1^*$,and there are enough free wavelengths to accommodate all the contending bursts. Their initial wavelengths will be shifted accordingly by the WCs.

deflected to path #1: The node's controller is in state, $q_1^* \leq q < q_2^*$.

deflected to path #2: The node's controller is in state, $q_2^* \leq q \leq \infty$.

Note that the threshold values q_1^* and q_2^* are set by taking into account the delay and blocking weights in equation (8).

III. QUEING MODEL ANALYSIS

In this section we analyze the queuing model of Figure 2. Recall that our objective is to minimize both jitter and blocking probability by routing bursts originating from a given source to a destination on a single path. To simply the model we will assume a single dispatcher queue and K path servers, each with service rates $\mu_j, j=\overline{1,K}$. Bursts arrive at a rate λ. Each server j represents an onward path with its own fixed QoS bounds i.e., jitter and blocking. When busy, the path exits this bound at a rate α_j and once exited, it tries to revert (restore) to this bound at a rate β_j. Deflecting bursts is dependent on the fixed queue thresholds q_1, q_2.

System states at any arbitrary time are [6]:

$$D_j(t)=\begin{cases}0, \text{original path in use or system idle}\\ 1, \text{ deflction route is busy}\\ 2, \text{deflection route is busy failing to meet QoS}\end{cases} \qquad j=\overline{1,K} \quad (10)$$

We can define a state space of the path servers as;

$$E_D=\left\{(d_1,d_2);\ \begin{vmatrix} d_j\in\langle 0,1,2\rangle, 0\le q\le q_1\\ d_1\in\langle 1,2\rangle, d_2\in\langle 0,1,2\rangle,\ q_1\le q\le q_2-1\\ d_1\in\langle 1,2\rangle, d_2\in\langle 0,1,2\rangle, (d_1,d_2)\ne(2,0),\ q_2\le q\le\infty\end{vmatrix}\right\} \quad (11)$$

From which we can define a state space as well as a random process respectively as:

$$E=\{x=(q,\mathbf{d}); q\in \mathrm{N}_o, \mathbf{d}=(d_1d_2)\in E_D\} \quad (12)$$

Under stationary conditions we also have;

$$\rho=\lambda\Big/\sum_{j=1}^{K}\beta_j\mu_j(\alpha_j+\beta_j)^{-1}\prec 1 \quad (13)$$

The utilization of each deflection path is:

$$U=1-\pi_{(0,0,0)} \quad (14)$$

where $\pi_{(0,0,0)}$ denotes an empty state space.

IV. ANALYSIS AND SIMULATION

In both our numerical as well as simulation performance analysis, we assumed the following:

$L-$ fixed data burst size $600MB$, and a BCP offset time of $0.4m\sec$. Each link has a capacity $C=6GBps$,

$\lambda-$ Burst generation rate of $120/\sec$. The network updating interval is fixed throughout the simulation runs. When a connection request arrives at a node a wavelength it is assigned along the least cost path:

$$c_o\le\frac{\alpha_i c_{1,2}+\beta_1 c_{1,1}}{\beta_1\mu_1}\le\frac{\alpha_2 c_{2,2}+\beta_1 c_{2,1}}{\beta_2\mu_2} \quad (15)$$

The evaluation is carried out on a multi-node-node network using OMNET++ (version 5.4.

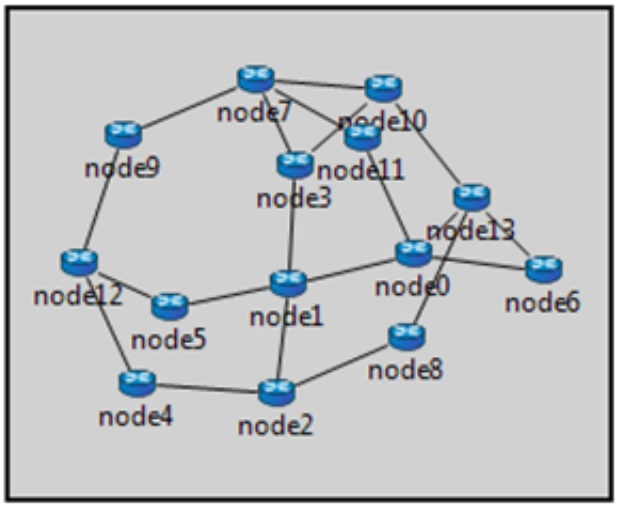


Figure 4. Network Model

In Figure 4 Node 0 is the source (s) whilst node 12 is the destination. Source routing using the random shortest path first algorithm is assumed.

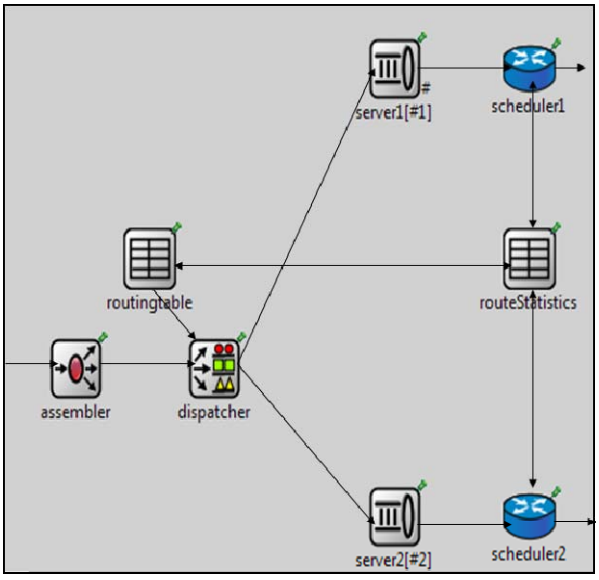


Figure 5. Edge node Configuration

We further make additional assumptions as follows:

- At the source node, all bursts are categorized according to QoS constraints e.g., blocking and delay.
- The various links constituting the network vary in length. They are also bidirectional with each fibre operating 16 wavelengths, 2 of which are dedicated for signaling purposes.
- Besides the original path, only two other deflection paths are available between this node and the destination.

This meaning the original path is preferred before deflection *path*#1 and *path*#2 respectively as it has the lowest cost.

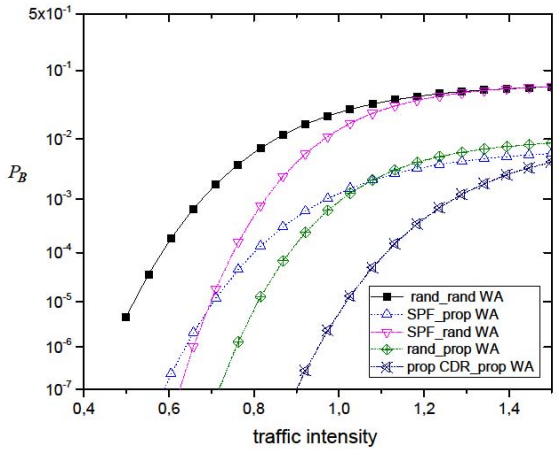


Figure 6. End to end P_B versus load

We first compare the performance of the proposed CDR as well as WA (prop CDR_prop WA), scheme with regards to loss probability (P_B).

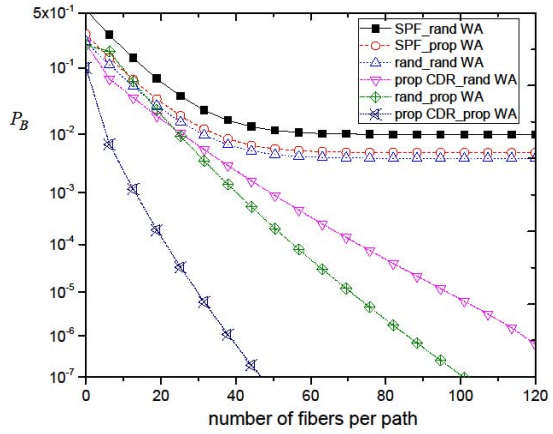


Figure 7. Loss probability as a function of number of fibres per path

In so doing we compare it with:

- CDR with random WA (prop CDR_rand WA,), in which the wavelengths are randomly assigned randomly.
- shortest path first together with random WA (SPF_rand WA).
- random path and random WA (rand_rand WA).
- SPF and proposed WA (SPF_prop WA.

Shown in Figure 6 is several plots of the P_B as a function of varying traffic loads. From this graph it is observed that the proposed CDR as well as proposed wavelength assignment (prop CDR_prop WA) outperforms the rest of the schemes.

Random routing coupled with the proposed WA (rand_prop WA) also shows fairly good performance as it tends to distribute traffic among the available routes. It is generally concluded that a combination of CDR and the proposed WA will reduce end to end blocking probabilities.

Figure 7 shows how the path blocking as a function of the aggregate number of wavelengths available on the path. The traffic load is maintained at 100%. An increase in the number of fibers per path results in reduced blocking. Noticeable is that the traffic is evenly distributed within the fibers and the traffic also uniformly spread, hence this leading to reduced blockings. W e also explore the effect of increasing the number of wavelengths on blocking. Once again, the Poisson arrival process is used in which each fiber's traffic load is set to 100%. The simulation scenario this time is repeated with three randomly chosen sets of ingress and egress node pairs.
As can be observed in Figure 8, the proposed scheme features much better when the number of wavelengths is increases, and at the same time the available resources are utilized uniformly and rationally.

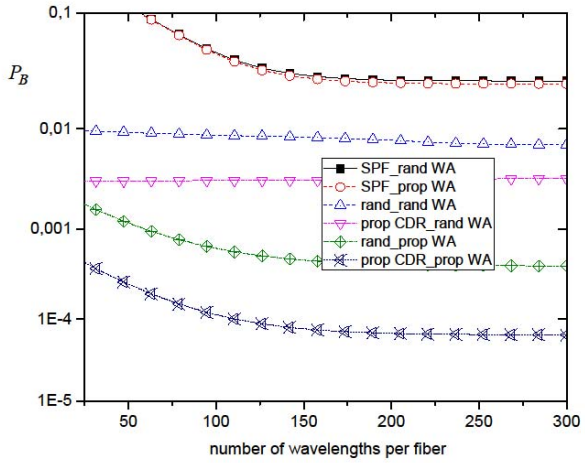


Figure 8. Loss probability as a function of number of wavelengths per fibre

We gradually increase the bursts arrival rate from 0 to more than 100% so that the controller queue is always above the q_2 threshold value and by so doing it is noted that deflection does reduce blocking, even though may propagate or trigger congestion/contentions in the deflected routes.. It may thus be necessary to regulate the volumes of deflected traffic. Figure 9 plots the performance of the various schemes as a function of total number of nodes traversed. The controlled scheme performs comparably better at high traffic volumes as it regulates the actual numbers reflected, e.g., some bursts are discarded.

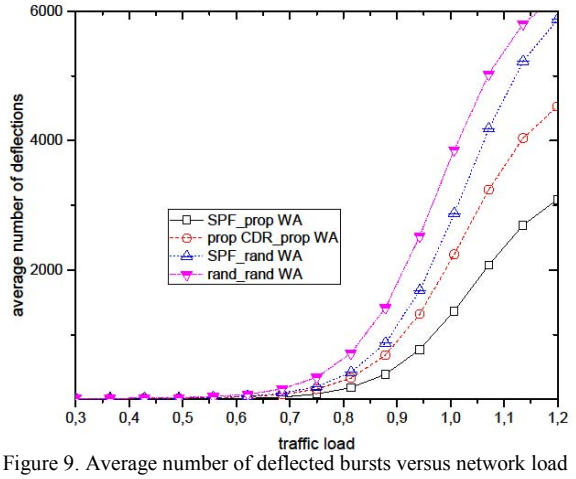


Figure 9. Average number of deflected bursts versus network load

End-to-end delays in the overall network are plotted as a function of the total number of nodes in Figure 10.

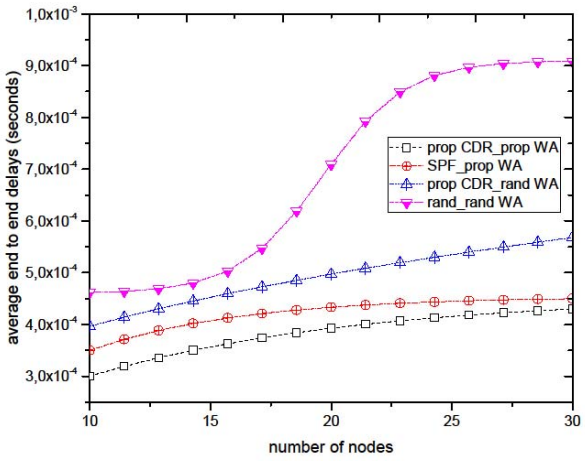


Figure 10. End to end delays versus number of nodes

In this case we compute the delays from the point of deflection. As seen from Figure 10, both the proposed scheme and SPF_prop WA more less feature alike. This is because fundamentally both opt for the shortest paths from the deflection point to ultimate destination egress node.

V. CONCLUSIONSS

In this paper, we proposed and described the Deflection Controlled Routing (CDR) based scheme that allows the deflection of bursts to alternate paths only after controller buffer preset thresholds are surpassed. The scheme couples with a proposed WA approach to significantly improve network performance especially in terms of delay and blocking probability QoS metrics. The proposed DRC scheme's performance is compared to other existing similar schemes or variants such as the ones discussed in [8] and [9]. Both analytical as well simulation evaluations were carried out. It is generally found out that the proposed CDR and WA scheme does significantly improve end-to end blocking as well as minimize end to end differential delays caused by bursts originating from the same source having to follow different paths. In that way jitter levels are minimized and its effects negligible.

REFERENCES

[1] K.Hirata, T. Matsuda and T. Takine, Dynamic Burst Discarding Scheme for DeflectionRrouting in Optical Burst Switching Networks", Optical Switching and Networking 4, 2007, pp106–120.

[2] S. Haeri, and L. Trajkovi´, "Intelligent Deflection Routing in Buffer-Less Networks", IEEE Transactions on Cybernetics, Volume 45, Number 2, February 2015.

[3] Ouveysi, I, Shu, F, Chen, W, Xiang, G, and Zukerman, M. (2010). Topology and routing optimization for congestion minimization in optical wireless networks. *Optical Switching and Networking*, Volume 7, Number 3, pp. 95-107. 2010.

[4] Klinkowski, M, Pedro, P, Pioro, M, Careglio, D, and Sole-Pareta, J, "Virtual Topology design in OBS Networks", IEEE 12th International conference, 27 June 2010.

[5] K. Christodoulopoulos, E. Varvarigos and K.Vlachos, "New Burst Assembly scheme based on the average packet delay and its Performance for TCP Traffic.", Optical Switching and Networking. Volume 4, Number 3, 2007.

[6] D. V. Efrosinin, M. P. Farkhadov and N. V. Stepanova, "A Study of a Controllable Queueing System with Unreliable Heterogeneous Servers", Automation and Remote Control, Volume. 79, Number. 2, 2018, pp. 265–285.

[7] A. I. Abd El-Rahman, S. I. Rabia, and H. M. H. Shalaby, "MAC Llayer Performance Enhancement Using Control Packet Buffering in Optical Burst-switched Networks," J. Lightwave Technology, Volume. 30, Number. 11, pp. 1578–1586, June 2012.

[8] P. Sakthivel, P. Krishna, "Multi-Path Routing and Wavelength Assignment (RWA) Algorithm for WDM Based Optical Networks", International Journal of Engineering Trends and Technology (IJETT) – Volume 10, Number 7 - April 2014.

[9] S. Li, M. Wang, E. W. M. Wong, V. Abramov, and M. Zukerman, "Bounds of the overflow priority classification for blocking probability approximation in OBS networks," J. Opt. Commun. Network., Volume 5, Number 4, pp. 378–393, Apr. 2013.

[10] S. Bregni, A. Caruso, and A. Pattavina, "Buffering-deflection tradeoffs in optical burst switching," Photon. Network. Communications., Volume 20, Number 2, pp. 193–200, Aug. 2010.

[11] E. Okly, N. Wada, S. Okamoto, N. Yamaka, Ken-ichi Sato, "Optical Networking Paradigm: Past, Recent Trends and Future Directions", IEICE Trans. Communications, Volume E100–B, Number 9 September 2017.

[12] Y. Ito, Y. Mori, H. Hasegawa, and K. Sato, "Optical networking utilizing virtual direct links," 42nd European Conference and Exhibition on Optical Communication (ECOC 2016), W.4.P1.SC6.3, Dusseldorf, Sept. 2016.

[13] K. Sato, "Optical Networking and Node Technologies for Creating Cost Effective Bandwidth Abundant Networks," Proc. OECC/PS 2016 (The 21st OptoElectronics and Communications Conference /International Conference on Photonics in Switching 2016), ThA1-2, July 2016, Niigata, Japan.

[14] Y. Uematsu S.Kamamura, H Date, H. Yamamoto, A. Fukuda, R. Hasashi, K. Koda, "Future Nationwide Optical Network Architecture for Higher Availability and Operability Using Transport SDN Technologies", IEICE Transactions in Communications, Volume.E101–B, Number 2, February 2018.

[15] A. Misawa S. Kataya,"Resource Management Architecture of Metro Aggregation Network for IoT Traffic", IEICE Transactions in Communications, Volume E101–B, Number 3, March 2018.

Advanced Multiplier Design and Implementation using Hancarlson Adder

E. Jagadeeswara Rao
Dept. of Electronics & Communication Engg.
Aditya College of Engg. & Tech., Surampalem,
AP, India
emandi.jagadeesh@gmail.com

T. Ramanjaneyulu
Dept. of Computer Science and Engg.
Aditya College of Engineering Surampalem
AP, India
ramanjaneyulu1214@gmil.com

K .Jayaram Kumar
Dept. of Electronics & Communication Engg.
Godavari Institute of Engg and Tech.(A),Rajahmundry
AP, India
jramworld@gmail.com

***Abstract*—Finite Impulse Response (FIR) filters, microprocessors and digital signal processors are the core system of multipliers. The Multiplier and Accumulator Unit (MAC) is the predominant block in a Digital Signal Processing (DSP) system. The objective of MAC is implementation of signal processing with high performance but multiplier most of the time occupies much area and become power consuming circuits. In this paper, a Modified Russian Peasant Multiplier (MRPM) using Hancarlson adder (HA) has been proposed. According to Russian Rules, a Divide and conquer technique is used in the multiplication process. But, in the perspective of digital design, only shifters and adders are used in the Russian Peasant Multiplier (RPM) to produce a Partial Product Generation (PPG). In this paper, we present an approach towards the reduction of delay in existing RPM by using HA, in the partial product reduction stage and proposed RPM with HA at Partial Product Addition (PPA). The proposed design is also compared to the RPM with Ripple Carry Adder (RCA), Carry Selector Adder (CSA) and 8-2 Adder Compressors (AC) in terms of propagation delay. The proposed design enhances speed of the system by 80.4% compared to the RPM using RCA, 81.7% compared to RPM using CSA and 77.5% compare to RPM with 8:2 adder compressors (AC's). The total operation is coded with Verilog Hardware Description language (HDL) using Model-Sim 6.3C, synthesized by using the Xilinx Integrated Software Environment (ISE) 14.7 design tool.**

Keywords—MRPM, RCA, CSA, 8:2 AC, PPG, HA

I. INTRODUCTION

Multiplication, which is an arithmetic operation, is a highly tiresome process. Multipliers are used as the main components in arithmetic, signal and image processors. Signal and image processing consists of functions like multiply, accumulate, convolution and filtering. The operation rate of a multiplier unit impacts the execution time of several processes. Multiplication takes more time compared to other operations in DSP algorithms, so critical delay path is calculated for a complete operation based on delay required for multiplication unit and it measures algorithm performance. The most widely used operations in computer arithmetic are addition and multiplication in case of approximate computing full-adder calls extensively analyzed for addition operation [1][2][3].

All DSP algorithms would need some form of multiplication and accumulation operations. MAC consists of adders, multipliers and accumulators. Usually DSP adders are RCA and CSA. Generally, multiplier multiplies the input values and passes the results to the adder, the adder then adds to the previous result of accumulator.

The organization of this paper is described as follows. Section II is a review for RPM with different schemes. The designs of an 8-2 AC are presented in Section III. Introduction Advanced adders are given in Section IV and existing RPM and proposed RPM are given in Section V and VI. Simulation results for multipliers with the HA is provided in Section VII and Section VIII concludes the manuscript.

II. LITERATURE SURVEY

Chang and Hsiao described the CSA that required a single RCA with zero carry-in and one sub-circuit such as a multiplexer. The design had a minimum transistor count and also an increase of 2-input in the NAND gate latency. The additional circuit is used to change the exact CSA with $C_{in}=1$. The design can reduced the transistor count by 29.25% and reduce the latency 5.9% for n=64 [7][16].

Gunasekaran and Manikandan developed a reconfigurable FIR filter using RPM which improved the addition operation in the MAC unit. Their design combined the CSA and the Sklansky Adder. It also reduced the area by 30.9% compared to conventional CSA and also improved the performance of the Carry generate (CG) block in CSA [14][17]. El-Guibaly modeled a fast and parallel MAC using the Modified Booth Multiplier (MBM). The dependence graph (DG) to describe the MAC hardware depends on MBM. The design of the Booth encoder is used by the carry-save technique. This proposed technique uses DG and a fast-pipelined model. The design process uses an accurate delay model for deep submicron CMOS technology [15][18].

III. COMPRESSORS

Compressors by far have been considered as the most efficient building blocks of a high-speed multiplier. It provides an advantage of accumulation of partial products at an expense of least possible power dissipation. Rather than entirely summoning partial products with the help of CSA/Ripple adder tree, a structure of compressors would complete the same task in much lesser time and also will simultaneously eradicate the problems of large power consumption and optimization of the area. This addition of partial products when done using conventional method of implementing full adders and half adders cannot account as much to lessening of delay associated with the critical path as when counter or compressors are used. The reason for the apparent preference of compressors over counters is the advantages it provides in terms of power, number of

transistors used, and the delay associated with the critical path (comprising of XORs mainly) [4]. The compressor design implemented in this paper prefers both Multiplexers (MUXs) and XORs.

The internal diagram of the 3-2 AC is illustrated in Fig. 1(a). The maximum delay of 3-2 AC is given by 2 XOR gates. The final sum S of the 3-2 AC is given in expression (1). The 3-2 AC can also be used as a full-adder (i.e. mux-based full-adder) when the input C is used as a carry input.

$$S = Sum + 2 * Carry \qquad (1)$$

The internal diagram of the 4-2 AC is illustrated in Fig.1 (b). The 4-2 AC has a reduced critical path compared to conventional adders since the maximum delay is given by 3 XOR gates. The 4-2 AC has five inputs (A, B, C, D, C_{in}), where C_{in} is the input carry, and three outputs (Sum, Carry and C_{out}). In this AC, the carry output C_{out} is independent of the input carry (C_{in}), making it possible to implement this structure with higher performance. The final sum S result of the 4-2 adder compressor is given in (2).

$$S = Sum + 2 * (C_{out} + Carry) \qquad (2)$$

The structure diagram of the 5-2 AC is presented in Fig. 1(c). The 5-2 AC maximum delay is given by 6 XOR gates. The final sum S of the 5-2 AC is given in (3).

$$S = Sum + 2* (C_{out1} + C_{out2} + Carry) \qquad (3)$$

The structure diagram of the 7-2 AC is presented in Fig. 1(d) [5]. The maximum delay is given by 10 XOR gates.

The final sum S of the 7-2 AC is given in (4).

$$S = Sum + 2 * (C_{out1} + C_{out2} + Carry\ ) \qquad (4)$$

In this paper, an 8-2 AC design using 3-2, 4-2, 5-2 and 7-2 AC. The structure diagram of the 8-2 AC is presented in Fig. 2(a,b,c,d) [6]. The final sum S of the 8-2 AC is given in (5).

$$S=Sum+2*(C_{out0} + C_{out1} + C_{out2} + C_{out3} + C_{out4} + Carry) \qquad (5)$$

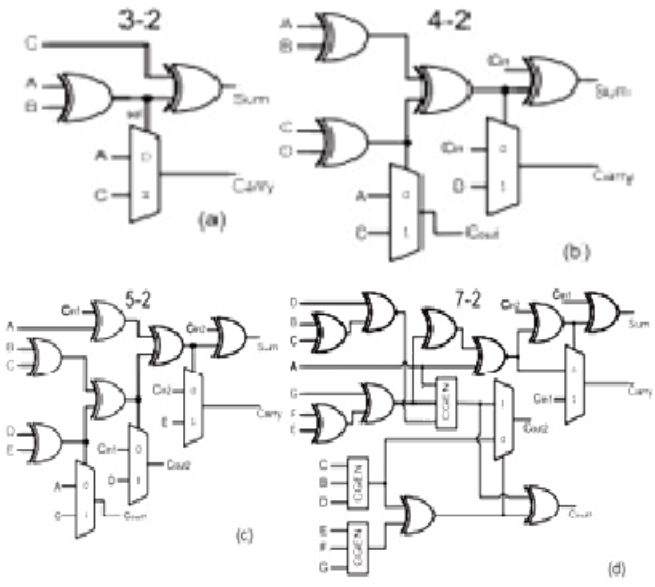


Fig. 1. AC structures: (a) 3-2; (b) 4-2; (c) 5-2; (d) 7-2 [5]

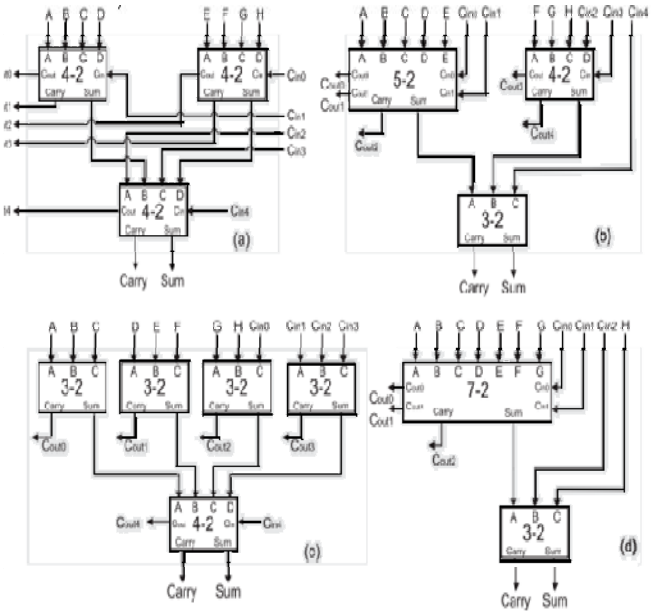


Fig. 2. the internal structure of 8-2 AC using: (a) Only 4-2 AC's; (b) Combination of 5-2, 4-2 and 3-2 AC's; (c) Combination of 3-2 and 4-2 AC's; (d) Combination of 7-2 and 3-2 AC's [6]

IV. Advanced Adders

A. 6-Bit Adder using 8-2 AC's

For any multiplication algorithm contains three steps but in this addition of partial products (PP's) is an important step to generate the final result. The multiplier performance depends on how fast PP's get added to obtain the final result. Many researchers can work in this area to achieve fast adders. The fundamental adder architecture is a RCA and further develops number of adders such as Carry Lock-ahead adder (CLA), CSA, and Carry save adder (CSAA) and Carry skip adder (CSKA) etc. In this RCA is well known for its regular structure and maximum delay because each step waits for the carry from the previous step.

CLAs have a minimum delay but maximum area associated with these adders. CSKA gives the more performance than RCA but it consists of extra hardware circuitry to skip the carry generated [8]. CSAA gives the further addition by reducing addition there are number of three into two. The major drawback CSAA consumes larger area [9]. Further carry select adder uses the two RCA and it does not wait for previous stage to execute. The CSA with higher bits exhibits excellent area and speed trade off compare with other adder architectures [10]. Many modifications can be done in CSAA for sacrificing its speed for area [11]. Fig. 3 represents an adder for adding eight 16-bit [12].

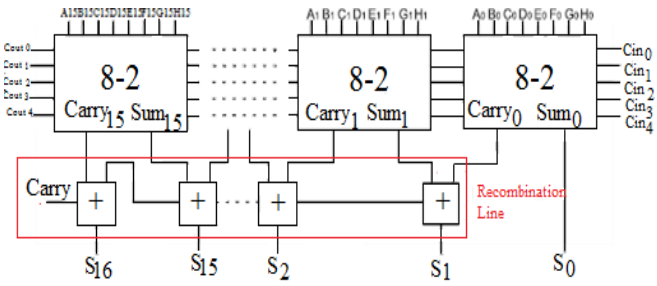


Fig. 3. 16-bit adder using 8-2 AC [12]

B. *Hancarlson Adder*

In the extension method to improve the performance of RPM using HA is used in the place of PP addition stage which resulted in increase in the speed of the RPM.

The formula for binary addition problem is generation of n-4, sum $s=s_{n-1}$, s_{n-2}.so n-bit augends $A=a_{n-1}$, a_{n-2}….. $a_0$and n-bit addend $B=b_{n-1},b_{n-2}$……b_0.we get c_i is the carry out of the i^{th} bit. The computation of whole piece s_i and carry c_i is:

$$s_i=a_i \oplus b \oplus c_{i-1} \quad (6)$$

$$c_i=a_ib_i+a_ic_{i-1}+b_ic_{i-1} \quad (7)$$

The stages of prefix addition to register are sum pre-processing, prefix processing and post processing. The generate g_i and propagate p_i signals are:

$$G_i=a_i \bullet b_i \quad (8)$$

$$P_i=a_i \oplus b_i \quad (9)$$

If $g_i=1$ means carry is generated at bit I and when $p_i=1$, a carry is propagated through bit i. The concept of generate and propagate extended to a block of contiguous bits, from bit k to I with necessary condition k<i.

$$g[i:k]=\begin{cases} gi & if\ i=k \\ g[i:j]+p[i:j]g[i:k] & otherwise \end{cases} \quad (10)$$

$$p[i:k]=\begin{cases} pi & if\ i=k \\ p[i:j]p[j:k] & otherwise \end{cases} \quad (11)$$

Where: i≥l≥j≥k

A carry is generated in block k-1 if condition g[i:k]meets and a carry is propagated through the block if condition meets p[i:k].The expression for any bit i the carry c_i is

$$C_i=g[i:0]+p[i:0]c_{-1} \quad (12)$$

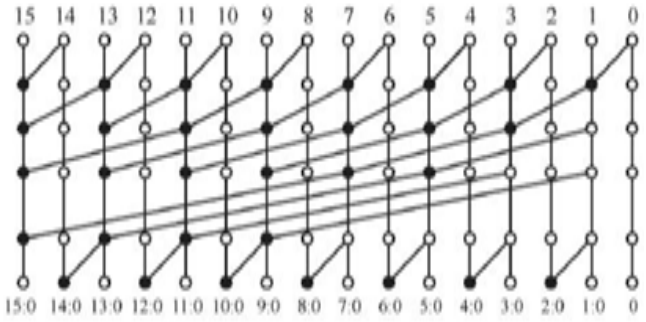


Fig. 4. HA Topology n=16 [13]

Where C_{-1} is the information convey of the n-bit adder. In the accompanying, for straight forwardness, we accept that $C_{i-1}=0$, so that above equation 4 follows as: $C_i=g[i:0]$

The prefix-preparing phase of the adder consists of block generate and propagate terms. So (g[i,k],p[i:k]) couples communicated with support of prefix operator shown below.

$$(g[i:k],p[i:k])=(g[i:j],p[i:j])\bullet(g[l:k],p[l:k])=(g[i:j]+p[i:j]g[l:k],p[i:j]p[l:k]) \quad (13)$$

V. EXISTING RPM

Existing RPM is designed to improve the hardware utilization of the circuit. The objective of VLSI system design is reducing power consumption, to improve speed and throughput of the system and minimize hardware complexity. Hence, the aim of proposed work is to reduce the delay and power consumption of multiplication. In general, multiplication function has three important steps [14]: Partial Product Generation (PPG), Wallace Tree Reduction (WTR) and Partial Product Addition (PPA).

The RPM reduces speed when partial product addition with RCA [7] and RPM with 16-bit AC has been illustrated in Fig. 5[15]. It gives n rows of partial products using only Multiplexers [8].

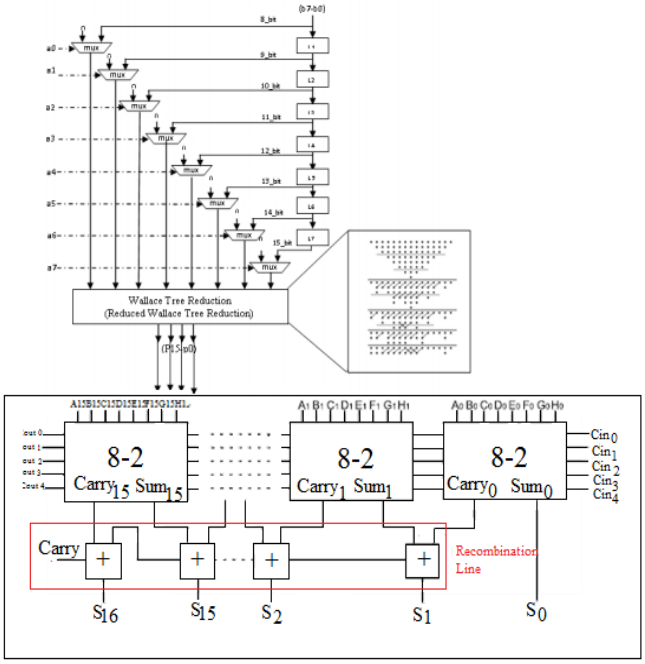


Fig. 5. Architecture of MRPM using 16-bit AC [15]

VI. PROPOSED RPM

The high speed MRPM depends on partial products generation and final sum. The proposed MRPM increase the speed by using HA at final sum stage as shown in Fig. 6. The Simulation Results of Proposed RPM and RPM with different adders at final sum stage discuss in Section VII.

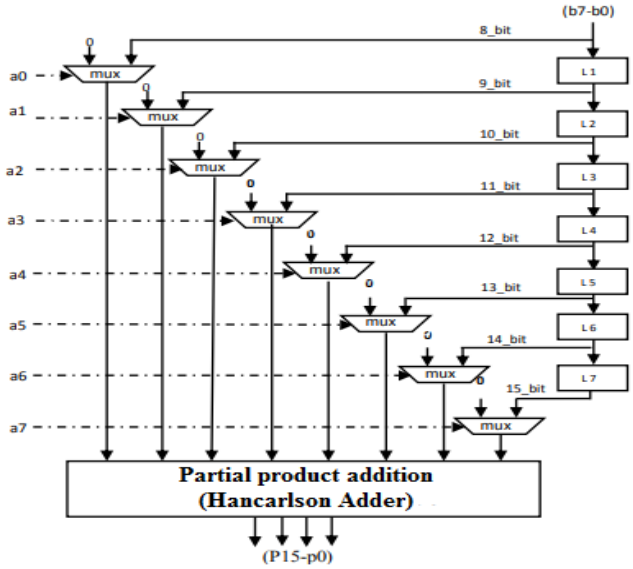


Fig. 6. Proposed MRPM using HA

VII. RESULTS AND DISCUSSION

The functional verification of existing RPM and proposed MRPM was done on Xilinx ISE. The targeted device is a Spartan-3e from the Spartan family. The grade speed of the design is set to -5. Table I gives the information about Slices, Lock-Up Tables (LUTs) and delay of RPM with different adders used at final sum stage.

TABLE I: COMPARISON EXISTING RPM AND PROPOSED MRPM OF 8X8 CONFIGURATIONS

Design	Slices	LUTs	Delay(ns)
Method-1	93	182	31.469
Method-2	88	161	18.646
Method-3	91	161	21.355
Method-4	79	142	16.556
Method-5	64	114	12.674
Method-6	70	123	15.551
Method-7	60	110	10.257

Method-1: RPM with CSA [13], Method-2: RPM with RCA [7], Method-3, Method-4, Method-5 and Method-6: MRPM with four 16-bit AC's [15] and Method-7: Proposed MRPM with HA. The design of four 16-bit AC's by using four 8-2 AC's [15]. Comparison of seven methods in terms of delay, slices and LUTs is shown in Fig. 7(a), (b) & (c). The delay of Method-7 is 10.257 ns, it is a less delay compares to remaining methods and large delay appears in Method-1 indicates in Fig. 7(a).

Less number of slices appears in Method-7 (60) and large number of slices appears in Method-1 & 3 (161) indicates in Fig. 7(b). Similarly, occupies less LUTs in Method-7 and large LUTs in Method-1 (182) indicates in Fig. 7(c).

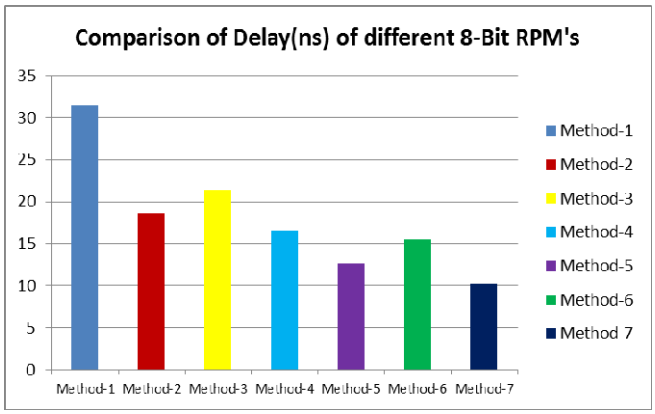


Fig. 7(a) Different RPMs comparison with delay (ns)

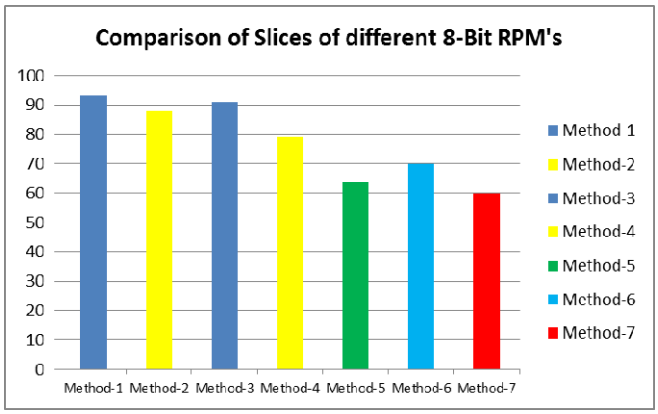


Fig. 7(b) Different RPMs comparison with slices

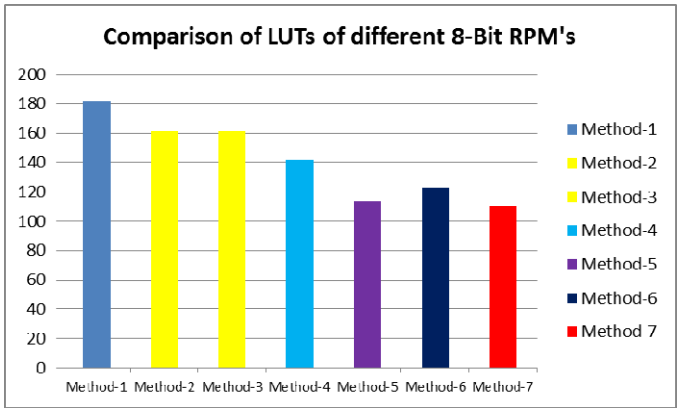


Fig. 7(c) Different RPMs comparison with LUTs

VIII. CONCLUSION

With results of the proposed design, it can be seen that the use of HA can enhance performance of the system significantly. The results and the comparisons were presented and clearly illustrating the advantages of the proposed design. It Enhances speed of the system by 80.4% as compared to the Existing RPM with RCA, 81.7% Existing RPM with CSA and 77.5 % compare to RPM with 8:2 AC's were achieved. The proposed MPRM unit offers 39.182% reduction in Slices, 0.277 % reduction in LUTs and 10.89 % reduction in delay than existing RPM. In future, Proposed MAC unit offers great advantage in minimizing the chip size for designing the communication standards.

REFERENCES

[1] V. Gupta, D. Mohapatra, S. P. Park, A. Raghunathan, and K. Roy, "IMPACT: IMPrecise adders for Low-Power Approximate Computing", *Proc. of the IEEE/ACM Int. Symposium on Low Power Electronics*, Japan, pp. 1-3, Aug. 2011.

[2] S. Cheemalavagu, P. Korkmaz, K. V. Palem, B. E. S. Akgul, and L. N. Chakrapani, "A Probabilistic CMOS Switch and its Realization by Exploiting Noise," *Proc. of the IFIP-VLSI SoC*, Perth, Australia, pp. 1-7, Oct. 2005.

[3] H. R. Mahdiani, A. Ahmadi, S. M. Fakhraie and C. Lucas, "Bio-Inspired Imprecise Computational Blocks for Efficient VLSI Implementation of Soft-Computing Applications", *IEEE Trans. on Circuits and Syst.*, vol. 57, pp. 850-862, Apr. 2010.

[4] V. G. Oklobdzija and D. Villeger, "Improving Multiplier Design by using Improved Column Compression Tree optimized Final Adder in CMOS Technology", *IEEE Trans. on Very Large-Scale Integration (VLSI) Syst.*, vol. 3, pp. 292-301, 1982.

[5] M. Rouholamini, O. Kavehie, A. P. Mirbaha, S. J. Jasbi, and K. Navi, “A new design for 7:2 compressors,” *Proc. of the IEEE/ACS Int. Conf. Comput. Syst. Appl. (AICCSA)*, Amman, Jordan, pp. 474-478, May 2007.

[6] J. S. Altermann, E. A. C. da Costa, and S. Bampi, “Fast Forward and Inverse Transforms for the H.264/AVC Standard using Hierarchical Adder Compressors,” *Proc. of the IEEE/IFIP Int. Conf. VLSI Syst. Chip (VLSI-SoC)*, Madrid, Spain, pp. 310–315, Sep. 2010.

[7] N. C. Sendhilkumar, "Design and Implementation of Power Efficient Modified Russian Peasant Multiplier using Ripple Carry Adder", *Int. Jour. of MC Square Scientific Research*, vol. 9, pp. 154-165, 2017.

[8] P. A. Guyot, B. Hochet, and J. Muller, "A Way to Build Efficient Carry Skip Adders," *IEEE Trans. on Computers*, vol. 36, pp. 1144-1152, 1987.

[9] M. Ortiz, F. Quiles, J. Hormigo, F. J. Jaime, J. Villalba, and E. L. Zapata, "Efficient Implementation of Carry-Save Adders in FPGAs", *Proc. of the 20th IEEE Int. Conf. on App. Specific Systems, Architectures and Processors (ASAP)*, Boston, pp. 207-210, Jul. 2009.

[10] R. Uma, V. Vijayan, M. Mohanapriya and S. Paul, "Area, Delay and Power Comparison of Adder Topologies," *Int. Jour. of VLSI and Communication Syst.*, vol. 3, pp. 153-168, 2012.

[11] B. Ramkumar and H. M. Kittur, "Low-power and area-efficient carry select adder," *IEEE Trans. on Very Large-Scale Integration (VLSl) Syst.*, vol. 20, pp. 371-375, 2012.

[12] B. Silveira, G. Paim, C. M. Dinizand and S. Bampi, "Power-Efficient Sum of Absolute Differences Hardware Architecture Using Adder Compressors for Integer Motion Estimation Design", *IEEE Trans. on Circuits and Syst.*, vol. 64, pp. 1-12, 2017.

[13] S. M. Sudhakar, K. P. Chidambaram and E. E. Swartzlander, “Hybrid Han-Carlson Adder”, *Int. Midwest Symposium on Circuits and Syst. (MWSCAS)*, Boise, Idaho, pp. 818-821, Aug. 2012.

[14] C. Uthaya Kumar and B. Justus Rabi, “Design and Implementation of Modified Russian Peasant Multiplier using MSQRTCSLA based FIR Filter", *Int. Jour. of Science and Tech.*, vol. 9, pp. 1-6, 2016.

[15] E. Jagadeeswara Rao and A. Rama Vasantha, “Design and implementation of high speed modified Russian peasant multiplier using 8-2 adder compressors”, *International Journal of Research in Electronics & Communication Engineering*, vol. 6, pp. 379-383, 2018.

[16] T. Y. Chang and M. J. Hsiao, "Carry-Select Adder using Single Ripple-Carry Adder," *Electronics Letters*, vol. 34, no. 22, pp. 2101-2103, 1998.

[17] K. Gunasekaran and M. Manikandan, "High Speed Reconfigurable FIR Filter using Russian Peasant Multiplier with Sklansky Adder", *Research Journal of Applied Sciences, Engineering and Technology*, 8(24), pp. 2451-2456, 2014.

[18] F. El-Guibaly, “Design and Analysis of Arbitration Protocols”, *IEEE Transactions on Computers*, vol. 38, issue 2, pp. 161-171, 1989.

A Survey of Undergraduate Project Topics in Computer Science at the University of Botswana

Gontlafetse Mosweunyane
Department of Computer Science
University of Botswana
Gaborone, Botswana
mosweuny@mopipi.ub.bw

Nkwebi Peace Motlogelwa
Department of Computer Science
University of Botswana
Gaborone, Botswana
motlogel@mopipi.ub.bw

Gabofetswe Malema
Department of Computer Science
University of Botswana
Gaborone, Botswana
malemag@mopipi.ub.bw

***Abstract*—Model curricula such as ACM/IEEE recommend that characteristics of students graduating in Computer Science should include project experience. In line with these recommendations, students in their fourth year of undergraduate study in the Computer Science department at the University of Botswana have to complete a major project as part of fulfilling the requirements for the award of a degree. The project gives students the opportunity to implement some of the theories and techniques learnt during the course of their programme of study. Students' choice of project topics therefore heavily depends on what the curriculum offers in terms of diversity of course content and techniques taught. This paper presents an analysis of 154 students' projects in the Computer Science department at the University of Botswana for the period 2013-2016 with a particular focus on the topic trends and issues of diversity and relevance to both the local and international context. The topics' list is categorized using the Agglomerative Hierarchical Clustering method. The analysis of titles and abstracts shows an inclination towards web-based systems and mobile applications that are relevant to the local context. Also, the titles and abstracts do not reveal core Computer Science areas such as algorithms, machine learning and security. Trending topics in Computing, which are also typically done as project topics, also seem not to feature. To reveal an effective categorization and assess comparability to international trends the rest of entire project document (body) needs to be analysed.**

***Keywords*—project, Computer Science, topics, curriculum, students, final-year, undergraduate**

I. Introduction

Model curricula such as ACM/IEEE [1] recommend that characteristics of students graduating in Computer Science should include project experience. Studies have shown that projects are beneficial in student' studies as well as to academic institutions and to industry as well [2], [3]. In line with these recommendations, students in their fourth year of undergraduate study in the Computer Science department at the University of Botswana (CSUB) have to complete a major project as part of fulfilling the requirements for the award of a degree. Four types of Undergraduate degrees are offered in CSUB: Bachelor of Science in Computer Science, Bachelor of Science in Computing with Finance, Bachelor of Science in Information Technology and Bachelor of Information Systems in Computer Information Systems. Computing courses offered across the four programmes include Algorithms, Databases, Mathematics, Programming, Operating Systems, Artificial Intelligence, Software Engineering and Networking and Security.

The project is currently run as a 4-credit, final-year, final-semester undergraduate course. It involves students carrying out a guided development project involving research and development on a topic of study, either initiated by the students themselves or suggested by a member of the teaching staff. Projects are largely influenced by the contents of the courses offered and taken during the student's programme of study, their interaction with local industry, or current technological trends. In the CSUB most of the topics are selected by students themselves as opposed to being suggested by an academic.

The project monitoring and assessment is overlooked by a project coordinator and involves submission of deliverables at specified stages to enable the department to monitor students' progress. The main aim of the project is to enable students to apply knowledge and skills acquired throughout their programme of study by carrying out a problem-solving application project. Such projects should challenge students by being integrative, require an evaluation of potential solutions and give students opportunities to develop their interpersonal communication skills [4]. The project also aims to enable students to demonstrate integration of the various knowledge areas taught in Computer Science, including skills such as making a clear problem-solving project statement, reviewing relevant literature, applying an appropriate methodology to a given problem, writing a quality technical project and technical project presentation. Evaluation at CSUB is done through assessment of the various outputs during the various stages of the project such as alpha and beta prototype presentations and reports, project management through interactions with supervisor and documenting the visits (weekly diary sheets), final system presentation, and the final report.

The project gives students the opportunity to improve their technical skills. Students usually implement a prototype and in the process learn and demonstrate their competences which are highly valued in the industry.

It also helps them improve their problem-solving skills through breaking down the problem (analysis) and providing the most effective way to solve the problem under their constraints. This process involves evaluation of methods, techniques and optimization of competing constraints.

The project also helps improve independent learning by giving them the opportunity to learn on their own. Some of the methods, techniques and technologies applied in their solution will be new to them, not having being taught in courses taken. This instills the spirit of lifelong learning which they will need to carry on in their professional careers. They also gain research skills; although most projects do not follow the scientific research process they do follow a similar process called the design process which students should be familiar with to produce an effective solution.

Lastly it helps improve communication and writing skills through reporting their findings and solutions in the form of a presentation and report.

The objective of this study is to analyse the type of projects students select and how they address issues of diversity, relevance and comparability to international 'standards'. From this study we determine if there is any bias towards certain types of projects and what could be the possible factors that influence selection of such project topics. The results of the study will help CSUB to retrospect on the diversity and relevance of the projects topics and aid to review course syllabi and approaches in guiding students' projects selection.

The rest of the paper is structured as follows. Section II presents the methodology used to carry out the analysis. Section III presents the discussion of agglomerative clustering of projects while Section IV concludes the paper and presents the recommendations.

II. Methodology

A. Data Collection

We collected a list of all the 154 projects done in the period 2013 to 2016 from current and previous project coordinators. This represents all the projects done during this period. The data consists of students' details (Id number, firstname, surname), programme of study, project title, supervisor and co-examiner details and approval details including recommendations and comments. Currently, there is no central repository for either the list of project titles or the outputs such as software and electronic copies of the reports. The student and project details are sourced from Microsoft Excel spreadsheets from various project coordinators, a role which rotates among academics almost every year in the department. A free Android OCR text scanner [5] was used to extract the paper-based abstracts/introductions to editable electronic text.

The data is compiled in a text file and anonymized, that is, details of the students are removed before being used in the next stages.

B. Categorization Using Word Cloud

We initially explore the topics by analyzing the data to deduce topic categories. The data consists of project titles and abstracts or a short introduction where the abstract was not written. We dig deeper into the data to find themes in the text to help visualize and summarize the topics done. For this we build semantic word clouds [6] from the titles and abstracts. Word clouds show the most important or frequent words in a document printed in a font scaled roughly proportionally to its importance, usually with visual properties such as font colour or word orientation varied [7]. It is meant to be visually scanned, with larger and more central words attracting attention [8]. While most tools provide a simple summarization of words without regard to relationships between them, some like in [6] and [9] use semantic linkage to place words that are related next to each other.

We build a semantic word cloud using the semantic word cloud visualization tool from the University of Arizona [10]. The tools take text as input and applies principles of information retrieval of term extraction, ranking and similarity computation to build the cloud. The semantic word cloud is shown in figure 1.

Fig. 1. Semantic word cloud showing project categories.

The word cloud shows a preliminary overview of topics done: mainly applications and management systems in health, communication systems and applications specific to the country (Botswana) context. While this provides a good overview of the topics done, the computing techniques or platform used is not clearly evident. Further analysis of either the documentation or software implementations done by students is necessary to determine the methods, tools and techniques used.

C. Categorization using Hierachies

The categorization of project topics also exhibit a hierarchical structure in the sense that we can have categories and subcategories. For example, we could have a web-based project that includes both networking and database components, where the later are subcategories. This categorization is well suited for a hierarchical clustering algorithm, for which there are algorithms and tools available to use for analysis and visualization.

We apply the agglomerative hierarchical clustering (AHC) algorithm, a process of assigning objects or documents to clusters by initially assigning each to its own cluster then repeatedly refining to merge pairs of clusters until an all-inclusive cluster is obtained [11]. The AHC algorithm builds a hierarchical tree using the bottom-up approach, building the tree from the bottom (leaves) to the top (roots). Using a number of approaches to compute similarity between two clusters, the most similar clusters are merged at each step.

Projects are initially manually allocated to some categories (a category being the subject area, technology or platform used, etc) and the distance matrix is developed. We build the distance matrix as a multi-category matrix which simply indicates whether a project topic belongs to a defined category or not (the former being 1 while the latter is 0). XLSTAT [12], a Microsoft Excel Add-In, was then used to apply the algorithm using default parameters. The metric used is the default Euclidean distance and the linkage criteria is Ward's method. Experimenting by varying the number of classes for truncation revealed an optimal class size for our dataset as 6 as the classes converged to manual clustering by domain experts. Figure 3 shows the dendogram showing optimal hierarchical categorization while figure 2 shows the categories related to $C1$–$C6$. The summary categorization is shown in figure 4.

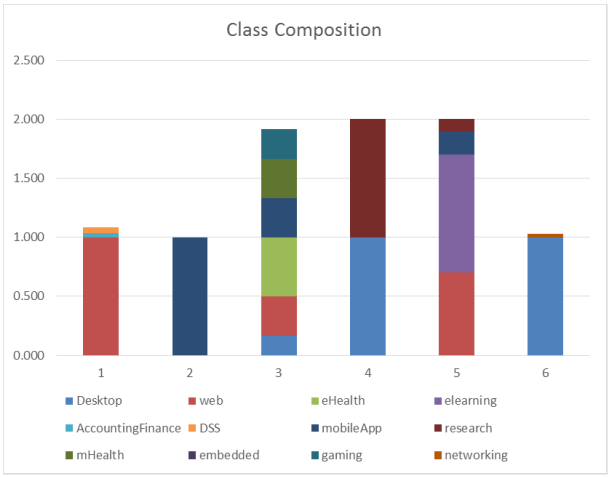


Fig. 2. Class Composition.

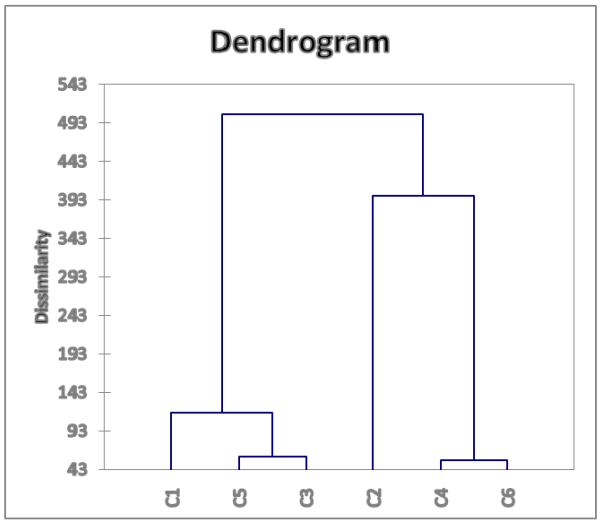


Fig. 3. Dendogram showing optimal hierarchical categorization.

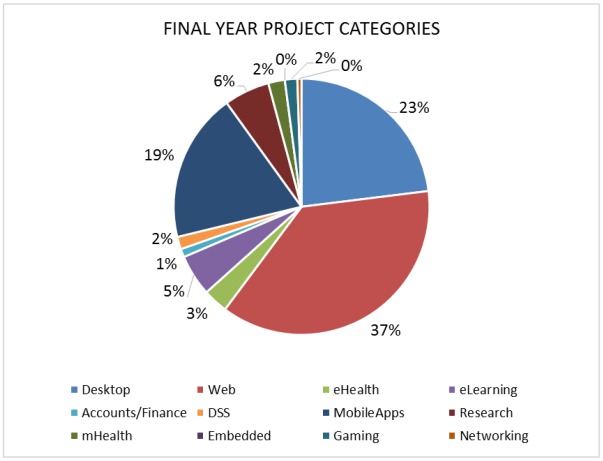


Fig. 4. Summary Categorization.

III. Discusion

The figures given in the previous sections define categorization of the project topics. Close to half (37%) of the final-year projects are web-based. This could be due to two factors. The popularity of web-based applications in student experiences such as course management systems, student records systems, library systems and social media may have influence on the popularity and affinity to web-based applications. The other reason might be the familiarity with the software development stack used for web-based projects. Students take a third year course on web technologies and applications. It will be more interesting to analyze implementing technologies such as the programming languages and database backend used in web-based projects.

Desktop applications make up 23% of the projects. This is the traditional category and it seems students are still interested in pursuing these. The other reason might be that these are supervisor-suggested projects that are taken by those students who could not come up with their own topics.

Mobile applications come third with 19% of projects in this category. Botswana has one of the highest mobile penetration in Africa [13]. In addition, the cost of smart phones and that of internet has dropped drastically. The lower uptake in mobile application projects cannot be attributed to students not owning smart phones or having difficulties accessing the internet. Rather it could be attributed to fear by students to take up a totaly new software development stack as there is no course at our University teaching mobile application development. Both android and windows mobile applications topics are being developed with no evidence of iOS and hybrid applications.

The remainder of the projects come from the gaming category. Although students like games a lot, it seems not many are interested in developing them.

The dendrogram in figure 3 shows the final progressive grouping of the final year project topics based on truncation into 6 classes. The first class is composed of all web projects,

a fraction of which is also accounting/finance or Decision Support System (DSS) applications. Class 2 is mobile applications, while class 3 is composed of desktop, web, and eHealth applications incorporating mobileApps, mHealth and gaming applications to some extent. Class 4 has desktop projects, part of which are research projects. Class 5 has mainly web, and eLearning projects, with a fraction being mobile Apps and research. Lastly class 6 is desktop projects, a fraction of which is netwoking projects.

The dendogram, though not providing unique clustering, could be cut at some level to reveal some partitioning that could be used to interpret the data structure and clusters [14]. The recommendation is to cut the dendogram where there is a large distance between two merged clusters, and since the within-cluster distance may be different for different clusters, it might be useful to cut it at different levels for each branch [14]. Therefore cutting the dendogram at 243 results is 3 clusters, the first composed of $C1, C3$ and $C5$ (all), the second being $C2$ (mobile Apps) only and the third being $C4$ and $C6$ (Desktop-research/networking). While the first and the last cluster make sense (a chuck of the projects are mobiles apps and a significant number are also desktop tools), the other, all-inclusive cluster seems to relate all the projects together and does not offer any significant advantage in categorization.

IV. Conclusions and Further Work

Applying the AHC algorithm to the projects lists using XLSTAT repeatedly and analyzing the resulting classes resulted in determination of the optimal number of classes as 6 as seen in figure 2. Cutting the resulting dendogram reveals 3 major clusters, two of which make sense. However, further investigation into cluster allocation is needed. An inspection of the project documentation submitted by students could be utilized to shed more light into how clustering might be refined. The distance matrix currently only indicates whether a project belongs to a category or not. Optimal clustering could be obtained by assigning values to distances between clusters as opposed to the current $0/1$.

We observe from the word cloud and AHC clustering that the majority of projects are simple applications addressing a need in the local context (for example a mobile app, an inventory or record management system or a website). The titles and abstracts do not reveal core Computer Science techniques and methodologies such as algorithms being applied to a specific problem. This is essential to achieve the objective of demonstrating and improving their technical skills. The majority of the topics only emphasize the platform (e.g. network, online, mobile) or application area (e.g. health, accident, hospital, booking).

From literature and surveying examples of final-year undergraduate development projects in Computer Science in other universities it can be seen that most projects involve a application or demonstration of a computing concept or area such as algorithm design and implemetation, development of a database system, distributed systems, artificial intelligence, networking and security and big data. Software development life cycle activities can then be applied in project development [15], [16]. The analysis of titles and abstracts at CSUB reveals a few of these areas (DSS, Embedded, Networking, routing, gaming). Most project seem to be development of artificats such as websites and management systems.

With regard to international trends it appears that newer and key Computer Science categories are not being explored using these projects. These include current highly trending topics in computing such as high performance computing, big data and artificial intelligence. In addition, students spend a lot of time on social media, but topics reveal little done from with regard to that category. While the survey has shed light into the various topics in CSUB final-year undergraduate students projects, there is need to explore further the tools, methods and techniques used through analysis of project submissions. To determine if teaching affects project topics, why and what influences such topic choices and how to fix challenges mentioned above further work will be carried out. The current survey, however, could serve to inform the department on the variety of projects done and serve as inital input for self-assessment and curriculum review. Our preliminary recommendation to CSUB is that there needs to be more guidance in selecting topics to engage students in core and trending areas in computing.

References

[1] Joint Task Force on Computing Curricula, Association for Computing Machinery (ACM) and IEEE Computer Society, *Computer Science Curricula 2013: Curriculum Guidelines for Undergraduate Degree Programs in Computer Science*. New York, NY, USA: ACM, 2013, 999133.

[2] J. López, B. Tormos, J. M. García, P. Olmeda, and C. Guardiola, "The importance of the final year project as a way of completing an education in engineering," in *International Conference in Engineering Education*, Valencia Spain, July 21-25 2003, pp. 1–4.

[3] A. C. Alves, R. M. Sousa, D. Carvalho, R. M. Lima, F. Moreira, C. P. Leão, L. C. Maia, D. Mesquita, and S. Fernandes, "Final year lean projects advantages for companies, students and academia," in *International Symposium on Project Approches in Engineering Education*, July 2014.

[4] T. S. M. T. Wook, N. M. Ali, M. Mohd, S. Saad, J. Salim, A. Yahya, S. M. Arif, L. Q. Zakaria, A. Ismail, S. F. M. Noor, and N. A. M. Zin, "Trends in the final year project for multimedia undergraduate programme: Supervisory experience," *Procedia-Social and Behavioral Sciences*, no. 59, pp. 399–405, 2012.

[5] Google Play , "Text scanner [ocr]." [Online]. Available: https://play.google.com/store/apps/details?id=com.peace.TextScanner &hl=en

[6] J. Xu, Y. Tao, and H. Lin, "Semantic word cloud generation based on word embeddings," *2016 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 239–243, 2016.

[7] M. Burch, S. Lohmann, F. Beck, N. Rodriguez, L. D. Silvestro, and D. Weiskopf, "Radcloud: Visualizing multiple texts with merged word clouds," in *Proceedings of the International Conference on Information Visualisation*, Paris, France, 16-18 July 2014.

[8] E. Schubert, A. Spitz, M. Weiler, J. Geiß, and M. Gertz, "Semantic word clouds with background corpus normalization and t-distributed stochastic neighbor embedding," *CoRR*, vol. abs/1708.03569, 2017.

[9] L. Barth, S. G. Kobourov, and S. Pupyrev, "Experimental comparison of semantic word clouds," in *Experimental Algorithms*, J. Gudmundsson and J. Katajainen, Eds. Cham: Springer International Publishing, 2014, pp. 247–258.

[10] University of Arizona. Semantic word cloud visualization. [Online]. Available: http://wordcloud.cs.arizona.edu/

[11] Y. Zhao, G. Karypis, and U. Fayyad, "Hierarchical clustering algorithms for document datasets," *Data Min. Knowl. Discov.*, vol. 10, no. 2, pp. 141–168, Mar. 2005. [Online]. Available: http://dx.doi.org/10.1007/s10618-005-0361-3

[12] XLSTAT 2017. Data analysis and statistical solution for microsoft excel. addinsoft. [Online]. Available: https://www.xlstat.com/en/

[13] Botswana Telecommunications Authority, "Bta annaul report 2011," Tech. Rep., 2011. [Online]. Available: http://www.bocra.org.bw/

[14] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *Trans. Neur. Netw.*, vol. 11, no. 3, pp. 586–600, May 2000. [Online]. Available: http://dx.doi.org/10.1109/72.846731

[15] V. Bouki, *Undergraduate computer science projects in UK: what is the point?* Thessaloniki, Greece: SEERC, 2007, pp. 176–184. [Online]. Available: http://www.seerc.org/ieeii2007/PDFs/p176-184.pdf

[16] P. Desai, , G. Joshi, and M. Vijayalaskhmi, "A novel approach to carrying out mini project in computer science & engineering," in *IEEE International Conference on Engineering Education: Innovative Practices and Future Trends (AICERA)*, Kottayam, 2012, pp. 1–4.

Services and Applications Security in IoT Enabled Networks

Philani Khumalo[1], Bakhe Nleya[2], A. Gomba[2] A Mutsvangwa[3]
[1,2]*Electronic Engineering Department, Steve Biko Campus, DUT*
South Africa
[1]massygomba@gmail.com, [2]bmnleya@ieee.org, [3]andrew.mutsvangwa@nwu.ac.za

***Abstract*—5G wireless together with optical backbone networks are expected to be the main pillars of the envisaged next /future generation networking (N/FGN) infrastructures. This is an impetus to practical realization of an IoT network that will support and ensure relatively higher bandwidth as well as enhanced quality of service (QoS) in both access and core network sections. The high-speed wireless links at the network peripherals will serve as a conducive platform for device-to-device (D2D) communication. D2D driven applications and services can only be effective as well as secure assuming the associated machine type communication devices (MTCDs) have been successfully verified and authenticated. Typically, D2D type services and applications involve the interaction of several MTCDs in a group. As such, secure and effective D2D group-based authentication and key agreement (AKA) protocols are necessary. They need to inherently achieve efficacy in maintaining the group key unlink-ability as well as generate minimal signalling overheads that otherwise may lead to network congestion. In this paper we detail a secure and efficient Group AKA (Gr-AKA) protocol for D2D communication. Its performance is compared to that of existing similar protocols and is found to comparably lower both computational as well as signalling overhead requirements. Overall the analysis shows that the Gr-AKA protocol improves performance in terms of fulfilling D2D communication's security requirements.**

Keywords—*5G network, group authentication, device-to-device (D2D), communication, security, privacy*

I. INTRODUCTION

The gradual shift from 3G/ 4G to 5G IoT enabled wireless as well as optical backbone networks will result in the provisioning of relatively higher bandwidth, lowered end-to-end latencies, massive device connectivity, reduced cost as well as consistent Quality of Experience (QoX). The shift also serves as a conducive platform for device-to-device (D2D) communication. Globally, D2D communication associated services and applications are steadily growing as billions of objects and devices are interconnected to form an Internet of Things (IoT) network [1].The combination of 5G wireless and optical backbones networks will enable capabilities of handling relatively larger mobile cellular data densities, support higher bandwidths to end users and serve vast numbers of objects and devices in comparison to current 3G/4G and optical networks.. This is illustratd in Figure 1. D2D communication will facilitate devices to interact directly without the aid of intermediary network elements [2]. In other words, it enables elements to communicate directly with each other without traversing fixed network infrastructural elements such as base stations or access points.

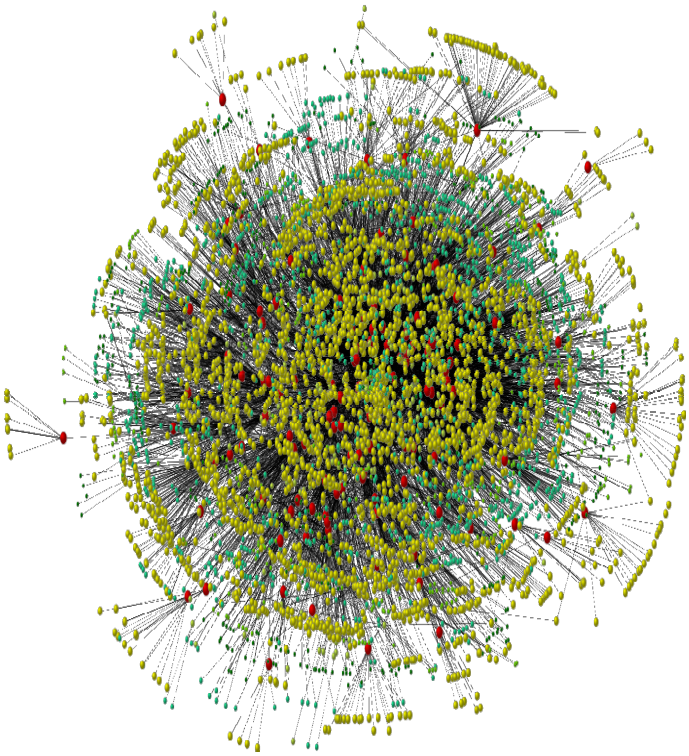


Figure 1. Devices in an IoT enabled Network

The main aim of D2D communication is to leverage the physical proximity of communicating objects as well as devices hence extending connectivity to sparse as well as remote environments [1]. D2D communication is primarily based in network infrastructural elements being involved in setting up a direct link between devices as well as the resources on which the communication will be facilitated. In classification terms, with *self-organized D2D communication*, the users themselves coordinate the setting up of the communication, whereas on the other hand with *network-assisted D2D communication*, the network infrastructural elements such as base stations (BS) assist direct data-transmission by means of control signaling and resource management [2]. D2D communication can further be distinguished as either,

out-*band D2D* in which users **communicate** over the unlicensed spectrum or In-*band D2D***,** in which the users utilize the licensed spectrum of the network operator.
In-band D2D communication can further be subdivided into (i) *Underlay in-band D2D* in which the D2D and users share the same frequency bands in order to increase the spectrum efficiency of the network and ii) *Overlay in-band D2D* in which D2D and users transmit over non-overlapping frequency bands [3].

D2D communication can be differentiated with Machine-to-Machine Communication (M2M) as the latter is essentially a paradigm similar to it. [4], M2M communication can be defined as "Data communication among machines or devices that do not require human mediation nor impose specific restrictions on communication ranges and is based on traditional networks such as 3G and LTE. It links communicating devices via routable core networks and M2M servers, even if the two devices are in proximity. It is essentially an application-oriented technology. D2D communication on the other hand assumes close proximity between devices and relies only on local device capabilities without centralized infrastructure support. It can be used for M2M communication to improve network performance and QoS [4].

Consequently, the next step is to facilitate interactions between humans and the devices/objects literally independent of human involvement [4]. Hence a need to facilitate traffic to traverse any network infrastructure. The success of D2D communication will facilitate key services and applications such as, health-care monitoring systems, cloud computing, smart transportation, intelligent tracking and tracing systems, smart cities as well as smart power grids. Fig.1 illustrates a few examples: -
Social entertaining services-sharing of multimedia files between trusted users in close proximity using social networking applications will be easily facilitated.
Public safety-D2D communication technology will enable areas that suffer natural disasters (e.g., hurricanes and earthquakes) still maintain connectivity with rescue personnel since D2D communication focuses on relaying information over the still available reliable short-range links.
Vehicular to vehicle (V2V) networks services-These will support the exchange of information between vehicles in close proximity in order to avoid catastrophic accidents and to improve traffic management.

Security and privacy now remain a key issue to be addressed when establishing D2D communication for the multitudes of MTC type devices (MTCDs) in an IoT enabled network. Prevailing security related protocols and technologies suffice for addressing most of the security issues in legacy networks but are not directly applicable to the IoT network as they are often memory and computing resources intensive. IoT objects and devices are often deployed and work in harsh, erratic and even intimidating environments, where they can easily be prone to various security breaches.

The various D2D based services and applications can only be effective as well as secure assuming the MTCDs have been successfully verified and as well as authenticated. Most D2D services and applications involve interaction of a group of MTCDs.

In this regard, several group-based AKA protocols continue to be explored for achieving effective authentication. Primarily they all must satisfy security requirements such as confidentiality, mutual authentication, integrity, privacy preservation and most importantly utilizing a common and single security (encryption) key during the communication sessions in the IoT network. Such protocols need to inherently achieve efficacy in maintaining the group key unlink-ability as well as generate minimal overheads that otherwise may lead to network congestion [6].

A: 3GGP MTC Arcitecture

An MTC architecture in a IoT network was proposed by the 3GPP committee [6] to facilitate the authentication between MTCDs and MTC users as illustrated in Figure 2.

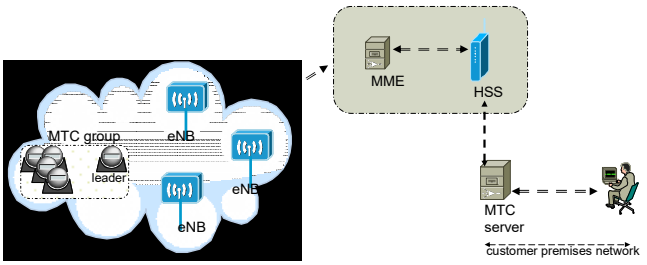


Figure 2. Communication Scenario [5]

The Mobile Management Entity (MME) as well as Home Subscriber Server (HSS) typically are located in network's pheriphery. The MTC server acts an as intermediary between MTC users and MTCDs. Users access it via an API. Secure communication between MTC serverand MTCDs requires prior authentication of the MTCDs by the network. This cannot be achieved using conventional protocols as they have short comings. Some of these short comings include, their tendency to generate excessive computational and signaling overheads as well as incapability to address key backward/foward secrecy (KBS/KFS) each time a device vacates or joins a group [7].

Generally, D2D devices do provide support for various features, normally geared towards optimizing the network resources available for use by applications. These are accessed as well as subscribed n the HSS. As such a single subscription can be utilized or shared by several devices. This will also include the associated security and access and control credentials associated with the subscription. Flexibility is granted in the form of subscribers being able to activate or deactivate some features whilst the subscription is valid. [8],[15].The 3GPP specification also addresses mechanisms for devices signaling congestion avoidance as well as traffic overload control. These are categorised as either soft or rigid.. With the soft apporoach, the service provider takes "soft" measures to try address any signalling congestion as well as traffic overload by minimising the number of attempts allowed per device.. With the rigid aproach associated devices are throttled altogether

Privacy and reliable mutual authentication and key negotiation are essential between the two communicating parties in D2D communication. In the absence of stringent privacy as well as security guarantees, data exchanged via the D2D communication links can easily be vulnerable to various attacks, e.g., eavesdrops, Man-in-the-Middle (MitM) and impersonation attacks. A user's identity can also be easily compromised [7].

The focus therefore is to secure D2D communication services such that user privacy and identity is never compromised. There is a tendency that a significant number of D2D applications and services will involve grouped devices and that they will be linked over several domains rather than within a single locality. In this kind of scenario, security issues are addressed taking into consideration that users are located under different domains, and that the D2D communication spans over several domains, operated by different operators, and with varying security policies.

Given the challenges and issues cited earlier, the Gr-AKA protocol for D2D communications (GAKA-D2D) that operates among multiple domains as well as operators. Its contributions are as follows: -

It ensures a group authentication mechanism that authenticates the group of participating MTCDs concurently.

- It preserves the security as well as privacy of the devices and at the same time maintains group key secrecy whenever a MTCD enters or vacates from it.
- For scalability as well as resource constraints, our proposal utilizes symmetric keys for authentication and authorization rather than asymmetric keys. As such it overcomes the security issuesof the network and generates relatively less signaling overheads.
- It satisfies key security requirements for D2D communication with moderate levels of both signaling as well as computational overhead.

II. PROPOSED PROTOCOL

The section commences by defining the system model as well as defining security assumptions. We then present the GAKA protocol. The system model is based on the generalized 3GPP MTC architecture as described in the previous section. We consider a conventional application scenario such as remote weather monitoring or crime surveillance. Initially, an MTC user registers with the local service provider for such a D2D communication service. This is followed by the network identifying a group of MTCDs ($MTCD_{grp_i}$) in the targeted area and initializing them. The group then designates a group leader ($grp_{i-leader}$) who will in turn negotiate both authentication and key establishment with the HSS/MME on behalf of the group. During this phase, the MTCDs and the HSS authenticate via the MME. Session keys are established between MTCDs and HSS for secure transmission of messages.

Session Key Compliance Stage: To ensure secured message exchanges between the MTCDs and the designated group leader, a session key is established among the group members. Because individual MTCDs may leave or exit the group; for each exit/or joining, key updating is necesary. A key generation center (KGC) communicates the updated information to all group member MTCDs.

MTCD join event: When a device joins a group, a new key is generated, so is the case when an existing member vacates the group.

MTCD exit event: A member can exit upon completing their task. In this case it must be prevented from accesing the group's resources, otherwise security is breached.. Hence the necessity to update current keys.

A. System Assumptions

We make the following assumptions:

- Considered is an applications or service in which several MTCDs together cooperate to form a group on one end and a single MTC user at the other end.
- The users are not necessary in their HOME location, and thus therefore prior registration (in case of roaming users) is necessary.
- The IoT service providers initially generate and agree on common system parameters as well as intra and inter operator agreements for D2D applications and services.

Asynchronous $(t;m;n)$ Group Authentication Scheme as proposed in [8] and further explored in [9] is utilized as the basis for carrying out group authentication.

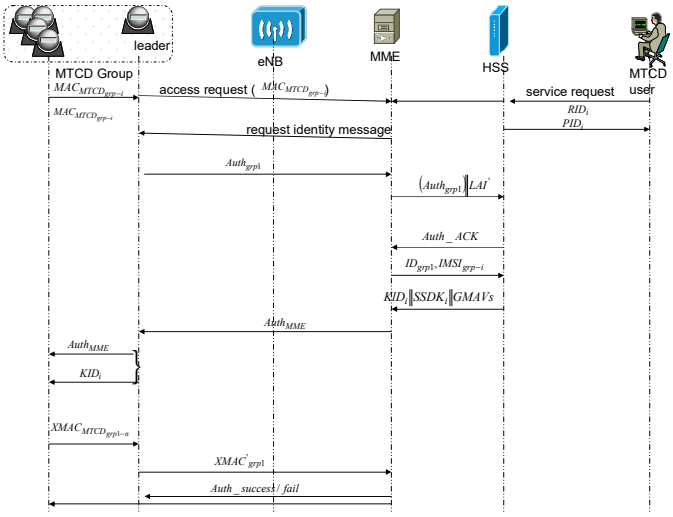


Figure 3. Sequence events for the proposed protocol

Asynchronous (t,m,n) group and authentication guarantees group authentication for m devices of a group with n members as well as being tolerant to t compromised tokens. In our protocol, it is considered that m has the same size of n, that is, all the members in a group are authenticated. Thus, it authenticates all the devices in a group simultaneously. The various sequence of events is summarised in Fig.ure 3. The detailed descriptions are narrated in the next section.

B. Session Request and Group Registration

A roaming MTC user registers for D2D services.

The MTC user (U_i) with a valid identifier (RID_i) performs user registration with the local HSS by furnishing his/her RID_i . If request is granted, the latter generates and issues a pseudonym ID (PID_i) to the user.

$$PID_i \overset{\text{def}}{=} (psedd, ExpiryTime) \quad (1)$$

The same RID_i will be used in the group initialization as well as key establishment process. The HSS also establishes and configures key parameters necessary for authenticating any formed MTC groups. Specifically, it generates a set of random numbers $\mathbf{R_z} \in \mathbf{Z_p^*}(z = 1,2,...i)$ and uses the set to compute a set of temporary identities $TID_{MTCD_{i-j}}$ to each $MTCD_{i-j}$ in a group:

$$TID_z = h_1(ID_{MTCD} \| \mathbf{R_z} * x) \quad (2)$$

where,
$h_1(.)$ is a secure hash function.
x is HSS's own secret key.

The HSS ultimately organizes the MTCD group into a binary tree [8]. Each node of the tree has a secret key that is known to each member MTCD. However, the secret keys of the nodes forming a path between a given MTCD and the root of the tree is not disclosed.
The HSS calculates a group key as follows:

$$GK_i = h_3(\sec_{i-1} \oplus \sec_{i-2} \oplus ..., \oplus \sec_{i-j} \oplus g * x) \quad (3)$$

where $g -$ is a random number, and $h_3(.)$ is a key generation function.

HSS further selects three hash functions; $h_1(.)$, $h_2(.)$, $h_3(.)$ which the key generation center (KGC) uses to generate an authentication message S to be used for group authentication. It also generates **k** tokens, all being a function of TID_{MTCD_i} to each device. These tokens must remain secret to any device outside the group. Finally the KGC computes and publishes the hash, function of S , $H(S)$ as well as hash function $H(.)$ that will be used to verify the validity of all MTCDs in the group.

C. MTCD Group Authentication and Key Agreement

This commences when a set of identified MTCDs within network coverage range request access so as part of a service/ application rendering. These are identified as a group ($MTCD_{grpi-j}$). We assume that within the group, the device with higher communication capability as well as battery reserve, will be designated as group leader($grp_{i-leader}$). The service provider then assigns a key (K_{grpi-j}) to each group member, as well as generating a group key which will be shared by both the MTCD group and HSS. The group key is used by individual MTCDs in the group for mutual authentication as well as privacy protection between MTCDs and service provider.

This is carried out mainly by the MTCDs' group leader and the HSS where the user is located. This is accomplished in the following sequence:

1. Each MTCD group member broadcasts a fresh temporary identifier $TID_{MTCD_{i-j}}$ and associated token $f(TID_{MTCD_{i-j}})$ to the group leader.

$$MTCD_{i-j} \rightarrow \lfloor TID_{MTCD_{i-j}}, f(TID_{MTCD_{i-j}} \rfloor \Rightarrow MTCD_{i-leader} \quad (4)$$

2. The group leader computes the Lagrange component vector for the group (LC_{MME}) using $TID_{MTCD_{i-j}}$ $f(TID_{MTCD_{i-j}})$ values received from the KGC.

$$MTCD_{i-j} \rightarrow LC_{grpi}$$

The general formula used is:

$$LC_{grp} = f(TID_{MTCD_{1-i}}) \prod_{q=1, q\neq j}^{\frac{n}{m}} \frac{-TID_{MTCD_{1-q}}}{TID_{MTCD_{i-j}} - TID_{MTCD_{i-q}}} \bmod p \quad (5)$$

This component is broadcast back to all group members. Each member uses it to verify whether all members are legitimate, by calculating the secret key S and comparing the result with $H(S)$ published by the KGC during registration phase.

3. The group leader further authenticates the group with the MME. In so doing, it first computes the group's $MACgrp_i$ and $Auth_{grp_i}$.

$$MAC_{grp_i} = h_2(GK \| ID_{grp_i} \| LAI \| S') \quad (6)$$

$$Auth_{grp_i} = (TID_{grp_i} \| MAC_{grp_i}) \quad (7)$$

$$MTCD_{grp_{i-leader}} \xrightarrow{Auth_{grp_i}, TID_{MTCD_{1-1}}, ...TID_{MTCD_{1-j}}} \rightarrow MME \quad (8)$$

4. The MME confirms with the corresponding HSS on whether the MTCD group is legitimate or not.

$$MME \rightarrow \underrightarrow{Auth_{grp_i}, LAI} \rightarrow HSS \quad (9)$$

5. Upon receipt of authentication verification request message from MME, the HSS authenticates the group by computing the group's MAC_{grp_i} using values received from the MME versus those it has in store.

$$MAC'_{grp_i} = h_2(GK \| ID_{grp_i} \| LAI \| S) \quad (10)$$

$MAC'_{grp_i} = MAC_{grp_i}$ implies successful authentication by the HSS, and the MTCD group leader will be informed accordingly.

HSS also further generates a temporary group key GTK for the MTCG Group.

$$GTK_{grp_i} = h_3(GK \| r_{HSS}) \quad (11)$$

Where, r_{HSS} is a random number. It also generates a token to MME that will enable the devices to authenticate the MME in future sessions

$$HSS \rightarrow \underrightarrow{f(ID_{MME})\|GTK_{grp_i}\|r_{HSS}} r \rightarrow MME \quad (12)$$

6. Upon receipt of messages from HSS, MME calculates its own Lagrange component LC_{MME} as well as $Auth_{MME}$ and broadcasts them to the MTCD group leader.

$$LC_{MME} = f(ID_{MME}) \prod_{q=1}^{\frac{n}{m}} \frac{-TID_{MTCD_{i-q}}}{ID_{MME} - TID_{MTCD_{i-q}}} \times \bmod p \quad (13)$$

$$Auth_{MME} = (LC_{MME}\|r_{MME} \oplus GTK\|r_{HSS}\|ID_{MME}) \quad (14)$$

Upon receiving $Auth_{MME}$, and encrypted KID_i the group leader broadcasts them to the rest of the group members.

7. Upon receiving the messages relayed from the MME each device updates its Lagrange component as follows:

$$LC\_new_{MTCD_{i-j}} = LC_{MTCD_{i-j}} * \frac{-ID_{MME}}{TID_{MTCD_{i-j}} - ID_{MME}} \quad (15)$$

Each device also uses the received r_{HSS} value to calculate GTK :

$$GTK_{grp_i} = h_3(GK\|r_{HSS}) \quad (16)$$

It also computes its integrity and cipher keys as well as $K_{asme}^{'MTCD_{grp_{i-j}}}$:

$$IK'_{grp_{i-j}} = h_4(ID_{grp_i}\|r_{HSS})K_{grp_{i-j}} \quad (17)$$

$$CK'_{grp_{i-j}} = h_5(ID_{grp_i}\|r_{HSS})K_{grp_{i-j}} \quad (18)$$

$$K_{asme}^{'MTCD_{grpi}} = KDF(GTK_{grpi}\|IK'_{grp_{i-j}}\|CK'_{grp_{i-ij}} \|ID_{grp_i}\|IMSI_{grp_{i-j}} \quad (19)$$

It further computes its own response value and sends it to the group leader.

$$XMAC_{MTCD}{}'_{grp_{i-j}} = h_1(ID_{grp_i}\|r_{HSS}\|IMSI_{grp_{i-j}})_{GTK_{grp_i}} \quad (20)$$

The group leader uses the response values from each of the group members to finally compute the group response.

$$XMAC_{grp_i} = h_1(XMAC_{MTCD_{grp1}} \oplus XMAC_{MTCD_{grp1-2}} \oplus \oplus,....XMAC_{MTCD_{grpn}})_{GRPK_1} \quad (21)$$

The Group leader finally passes the group response ($XMAC_{grp}$) to the MME for final authentication of each MTCD .

D. MTCD Joining or Exiting

In the event that a MTCD joins or vacates an already authenticated group, the secret S must be updated to avoid the old member to continue knowing the secret and to avoid new members discovering and exploiting previous secret values S. As illustrated in Fig. 4, when a MTCD joins, a new group key is generated:

$$GK_i^{'} = h_3(GK \oplus \sec_{i-j}) \quad (22)$$

Where $\sec_{i-j}$ is the secret value of the node to which the new MTCD is located. Likewise, HSS generates a new value for S as follows:

$$S_{new} = S + \delta S \,, \quad (23)$$

Where δS is a random value of S generated each time a member joins or exits.

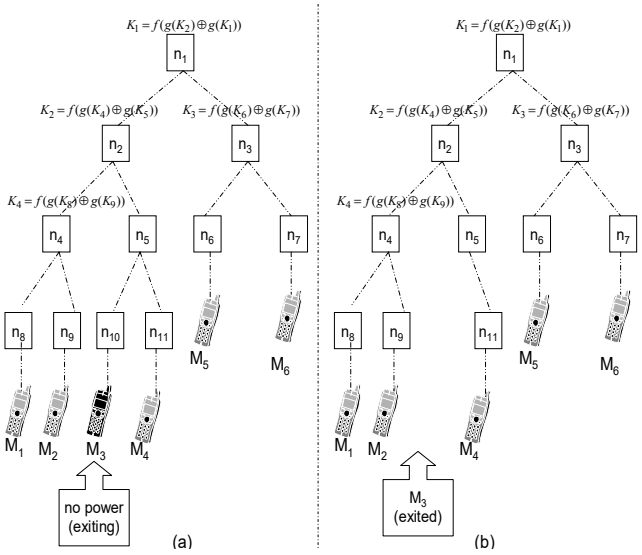


Figure 4. Example MTCD join/exit event tree

When a member exits, a new group key is computed according to:

$$GK_i^{''} = GK \oplus \sec_{i-j} \quad (24)$$

III. ANALYSIS AND PEFORMANCE EVALUATION

In this section we provide a general security analysis of the Gr-AKA protocol . Firstly we discuss its general security capabilities. We then go on to analyse metrics such as the amount of signalling overheads generated and exchanged bandwidth requirements, computational complexity as well as overall efficiency. We test some of the aspects of the Gr-AKA protocol using the AVISPA tool [10], [11].

A. Security Analysis

User's Privacy: At registration phase, **t**he MTC User's identity is mapped to a pseudonym ID (PID_i) and thereafter the latter is used for authentication purposes rather than the real User's name.. In this way the user's real identity is concealed hence privacy is guaranteed .

Mutual authentication: The Gr-AKA protocol provides robust mutual authentications between $User\_HSS$, as well as among the individual $MTCD_{grp_i}$ members. HSS authenticates the user by way of verifying MAC values computed using the User's credentials such as RID and PID. To authenticate HSS, the User checks the received MAC'' from the MME and if they both match with the $XMAC'$, then both MME and HSS are authenticated.

Similarly, HSS verifies and authenticates the $MTCD$ group by verifying ttheir their Lagrange components. Each member then uses these Lagrange components to compute the secret S and compares it with the same value that was sent from the KGC.

Backward /Forward Key Secrecy . With the Gr-AKA protocol the group key (GK) is updated and changed each time a device leaves or joins the group .When a device joins the group, HSS is compelled to broadcast its secret node, thus a new GK is computed using equation (22).
Similarly when a device exits, the remaining devices are compelled to update their GK using equation (24).

Attack resistivity: The channel between the MTC user and MTCD group is open to various attacks. To safeguard against replay attacks, time-stamped key hint messages are periodically exchanged between the two parties. A hacker who successfully intercepts the key hints exchange will not be able to replay a message for the next key hint exchange message because of the time stamping.

MiTM attack: The channel between the MME and HSS is assumed to be secure (in terms of integrity, confidentiality, and entity authentication), and only the channel between $MTCD_{grp_{i-j}}$ and MME may be vulnerable to MiTM attacks. However in the proposed protocol, the use of Shamir's secret sharing [9], together with the Lagrange component, makes it extremely difficult to recover the secret token. Furthermore, the group's ID is secret thus further making it difficult for attackers to generate or verify the MAC_{grp_i}.

B. Performance Analysis

The protocol is analysed in terms of its general security capabilities, computational demands/complexity as well as signalling overheads. The main security aspects of the protocol is tested using the Automated Validation of the Internet Security Protocols and Applications (AVISPA) tool[11]. We first created the model as in Figure 2, and also specifying the basic roles. We were able to verify that it can guarantee the privacy of a generated session key, as well as general authentication between MME and HSS.

To evaluate the total computational overheads, we compare the protocol to similar proposed protocols such as PPAKA-HMAC [13], G-AKA[13], and GBS-AKA [14]. In our analysis, we assume that overall there are n $MTCDs$ and each can have up to m members each. The following cryptographic computational times are utilised; maptopoint hash operation ($T_{mtp} = 0.07ms$), $MTCD$ Lagrange component computational time ($T_{L-MTCD} = 0.06ms$), HSS Lagrange computational time ($T_{L-HSS} = 0.04ms$), , multiplication over an elliptical curving ($T_{mul} = 0.6ms$), pairing ($T_{pair} = 4.5ms$), hash operation ($T_{hash} = 0.07ms$), symmetric ciphering/deciphering ($T_{aes} = 0.16ms$),. Table 1, summaries the computational overheads of the 4 protocols.

TABLE I. COMPUTATION COMPLEXITY OF GROUP PROTOCOLS [10]

AKA Protocol	Computational overhead		
	MTC Devices	Network	Total (ms)
PPAKA-HMAC [12]	$3T_{hash}*n+(T_{hash})*m$	$2T_{hash}*n+(T_{hash})*m$	$5T_{hash}*n+(2T_{hash})*m$ $2*Rand+(n+10)$ $*\exp(5+n)$ $*H_{mac}+(2n-1)$ $*Mul+1*Hash$
G-AKA[13]	$4T_{hash}+(4T_{hash})(n-1)$	$3(T_{hash})*n*(2T_{has}h)*m$	$7T_{hash}*n+(2T_{hash})*m$
GBS-AKA [14]	$(T_{\mathrm{mod}}+2T_{hash})*n+2T_{hash}*m$	$(T_{\mathrm{mod}}+2T_{hash})*n$ $+(6T_{hash}+T_{aes})*m$	$(2T_{\mathrm{mod}}+4T_{hash})*n$ $+(8T_{hash}+T_{aes})*m$
GR-AKA (proposed)	$(4T_{hash}+2T_{aes})*n$ $+(2T_{hash})*m$	$(3T_{hash}+2T_{aes})*n$ $+(2T_{hash})*m$	$(7T_{hash}+4T_{aes})*n$ $+(4T_{hash})*m$

We explored execution key generation time as a function of key size in bits. The key size is varied from 2^1 to about 2^{13} bits. In Figure.5, it is observed that as the key size is increased, so does the key generation time. However, increasing the key length makes it more secure. We thus chose to fix the key size to 12 bits (2^{10}), which is a 7 seconds delay. This is not so much a hindrance as key generation is a once off operation during initialization.

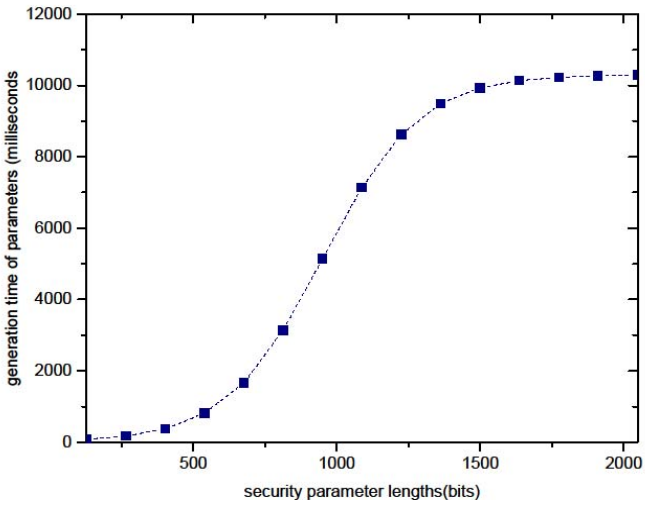


Figure 5. Overall protocol key generation time versus size

We further explore the proposed protocol's execution time and compare it with that of the PPAKA-HMAC [13], G-AKA[13],.GBS-AKA [14].

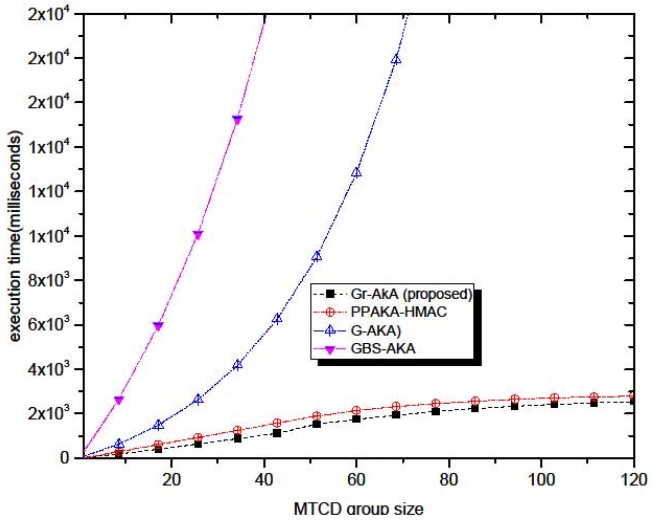


Figure 6. Execution time comparisons

The execution time more less increases linearly with increase in MTCD group size for the proposed protocol as well as PPAKA-HMAC[13]. However, we see an exponential increase in execution times with the other two protocols as clearly indicated by the graph of Figure. 6. We also analyse the overall magnitude signalling (communication) overheads of the proposed protocol. Overall the total signalling bits are computed from all the messages exchanged during the authentication process. Figure. 7 shows plots of total signalling (communication) overhead as a function of the number of MTCD groups, each comprising 5 members. Overall, the proposed protocol together with the PPAKA-HMAC generate more or less the same levels of signalling data, and not so excessive to cause congestion in the signalling channels.

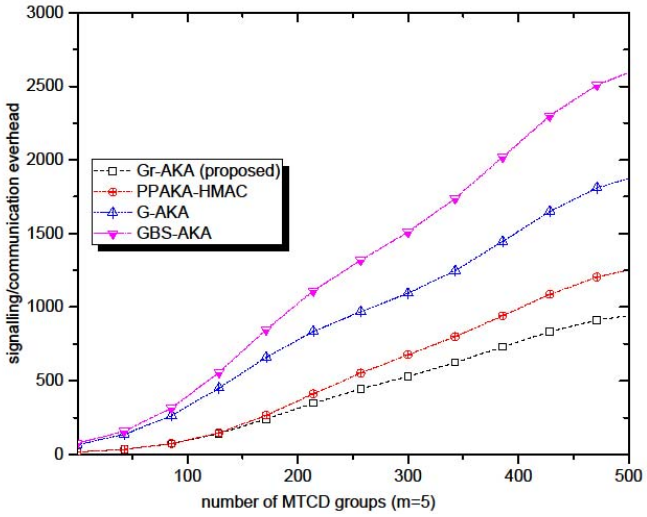


Figure 7. Signalling overhead

IV. CONCLUSION

We proposed the Gr-AKA protocol for enhancing security in D2D communication. It incorporates a secure group authentication mechanism that ensures MTCDs are authenticated as a group. As such, it preserves the privacy of the MTCDs and at the same time maintains the group's secrecy whenever a group member joins or vacates from it. The protocol is also designed to overcome the security problems of the network as well as the generation of relatively lesser overhead compared to other existing group-based AKA protocols.

REFERENCES

[1] E. Hossain and M. Hasan, "5G cellular: key enabling technologies and research challenges," IEEE Instrumentation & Measurement Magazine, vol. 18, no. 3, pp. 11–21, 2015.

[2] A. Michalas, V. A. Oleshchuk, N. Komninos, and N. R. Prasad,"Privacy-preserving scheme for mobile ad-hoc networks," in 2011 IEEE Symposium on Computers and Communications (ISCC), June 2011, pp. 752–757.

[3] G. Fodor, E. Dahlman, G. Mildh, S. Parkvall, N. Reider, G. Miklos, and Z. Turanyi, "Design aspects of network assisted device-to-device communications," IEEE Communications Magazine, vol. 50, no. 3, 2012.

[4] D. Feng, L. Lu, Y. Yuan-Wu, G. Y. Li, G. Feng, and S. Li, "Device to-Device Communications Underlaying Cellular Networks," IEEE Transactions on Communications, vol. 61, no. 8, pp. 3541–3551, 2013.

[5] Technical Specification Group Services and System Aspects; 3GPP System Architecture Evolution (SAE); Security Aspects of Non-3GPP Accesses (Release 11), document 3GPP TS 33.402 V11.4.0, 3GPP, Jun. 2012.

[6] M. Haus, M.Waqas, A. Y. Ding, Y. Li, S. Tarkoma, and J. Ott, "Security and Privacy in Device-to-Device (D2D) Communication: A Review," IEEE Communications Surveys & Tutorials, vol. 19, no. 2, pp. 1054–1079, 2017.

[7] C. Lai, H. Li, R. Lu, and X. S. Shen, ''SE-AKA: A secure and efficient group authentication and key agreement protocol for LTE networks,'' *Comput. Netw.*, vol. 57, no. 17, pp. 3492–3510, 2013.

[8] Harn, L. Group Authentication. *IEEE Transactions on Computers*, v. 62, n. 9, p. 1893-1898, 2012.

[9] Shamir, A. How to share a secret. *Communications,* pp. 612-613, 1979.

[10] A. Armando, "The AVISPA Tool for the Automated Validation of Internet Security Protocols and Applications", In Computer Aided Verification. CAV 2005. Lecture Notes in Computer Science, vol 3576. Springer, Berlin, Heidelberg.

[11] http://www.irisa.fr/lande/genet/span.

[12] M. Wang and Z.Yan,"Privacy-Preserving Authentication and Key Agremment Protocvols for D2D Group Communications", IEEE Transactions on Industrial Informatics, vol 17, 2017.

[13] Y. W. Chen, J. T. Wang, K. H. Chi, and C. C. Tseng, "Group-based authentication and key agreement," *Wireless Personal. Commun.*, vol. 62, no. 4, pp. 965-979, 2012.

[14] J. Yao, T. Wang, M. Chen, L. Wang, and G. Chen, "GBS-AKA: Groupbased secure authentication and key agreement for M2M in 4G network," in *Proc. IEEE Int. Conf. Cloud Computing. Reearchs. Innov. (ICCCRI)*, May 2016, pp. 42-48.

[15] G. Singh, D. D. Shrimankar, "Dynamic Group Based Efficient Access Authentication, and Key Agreement Protocol for MTC in LTE-A", Wireless Networks, Personal Communalizations, Published online, 13 April, 2018.

[16] 3GPP, "Service Requirements for Machine-Type Communications," TS 22.368 V10.1.0, June 2010.

A Review of Automated Detection Methods for Cyberbullying

Thabo Mahlangu , Chunling Tu and Pius Owolawi
Department of Computer Systems Engineering
Tshwane University of Technology
Gauteng, Pretoria, RSA
thabomahlangu44@gmail.com

***Abstract*—As we see the cyberspace evolve we also see a directly proportional growth of the people using the cyberspace for communication. As a result, the misuse of the cyberspace has given rise to negative issues such as cyberbullying, which is a form of harassing other people using information technology in a deliberate and continual manner. The detection and prevention of cyberbullying becomes critical for safe and health social media platforms. In this paper, a review of the cyberbullying content in Internet, the categories of cyberbullying, data sources containing cyberbullying data for research, and machine learning techniques for cyberbullying detection are overviewed. The main challenges of the cyberbullying detection are demonstrated, including the lack of multimedia content-based detection and availability of public accessible dataset. Suggestions are provided as the conclusion of the overview.**

Keywords—cyberbullying, cybersecurity, machine learning, SVM

I. INTRODUCTION

Social media platforms have become great platforms for connecting people from all over the world. However as we see these online social platforms gain more popularity on the cyberspace, we also see people take advantage of them to do great things and others conduct a variety of malicious activities on them. Cyberbullying is one of these negative issues develops with the social media platforms. Cyberbullying can be considered as a way of misuse with information technology to harass other people in an intentional and repeated manner. Most cyberbullying occur in a form of posting messages, tweets, rumours and sharing of embarrassing or private pictures and videos on social media platforms [1].

Bullies have adopted the platforms to conduct bullying activities behind the computer or mobile phone through the internet. The study performed by Symantec showed that about 25% of parents reported that their children have been involved in cyberbullying incidents [2]. Most works have been done in cyberbullying were focused on the impacts on people and how can a victim deal with it psychologically. Few work has been done on detecting the cyberbullying and preventing them before they take place using machine learning methods. This paper overviews the popularity of cyberbullying in social media platforms, the categories of cyberbullying and approaches to detect cyberbullying on social network by taking advantage of the existing machine learning practices.

The remainder of this paper is organized as follows. Section II presents the cyberbullying background and the types, Section III presents the various data sources for cyberbullying research, Section V presents a review of related works on cyberbullying detection, and Section VI concludes the paper

II. CYBERBULLYING IN SOCIAL MEDIA PLATFORMS

A. The Definition of Cyberbullying

This section focuses on the understanding the current knowledge about cyberbullying base on literatures. In the existing literatures, cyberbullying has been defined in different ways, which are helpful to gain understanding of the term in various fields.

In most common recognition, cyberbullying encompasses sharing or sending malicious messages or comments including pictures via innumerable communicating platforms such as SMS, emails, social network etc. The young kids and youngsters are the common targets of cyberbullying, because they are more vigorous to using novel technology, such as the internet [3]. Cyberbullying was defined as a type of harassment where one person is rude to the other, having to insult or offending another person [4]. Cyberbullying also means the use of technology to harm or harass other people in considered, repeated, and unfriendly manner. Cyberbullying differs itself to old-style bullying by happening 24 hours of the day and seven days a week. It occurs in innumerable forms such as text messages, spreading gossips, posting messages and sharing embarrassing images and videos on social media [1]. Cyberbullying is not so different from traditional bullying except it occurs on the cyberspace where a message can be sent directly to the victim or posted on public forums where everyone on that platform can see. These cyberbullies can be anyone sometimes are peers that share the same class or someone totally unknown by the victim, but most of the time they are familiar with their targets [5].

B. The types of Cybebullying

There are several types of cyberbullying existing in modern communicating platforms, for instance, types of cyberbullying were reported in [6],[21],[22],[7],[30] [31]and [37] as follows.

1) Flaming: beginning an online fight.

2) Harassment: the victim is frequently sent insulting and rude messages,this is the most common one since most litteratures are trying to address this specific type of cyberbullying.

3) Cyberstalking: The victim is sent intimidating or offensive messages. It creates fear to the victim.

4) Masquerade: the bully pretends to be someone who they are not.

5) Trolling: intentionally upsetting others on social networks by posting controversial comments.

6) Denigration: posting negative gossips about another person

7) Outing: posting private information about another person in public forums.

8) Exclusion: excluding someone from a social group.

9) Catfishing: when someone steal another person's identity online and creating a fake profile to deceive others

10) Dissing: is when someone post information or imagines about another person with the intention of damaging their reputation or popularity

11) Trickery: is when someone trick you to trust them so that you can share with them your secretes or personal private information that they can use to share it with the public online

12) Fraping: when someone use your social media account and pretend they are the owner of the account and then start posting inappropriate things to trick others thinking that you posted the content

C. *The Impact of Cybebullying*

This section explores the impact of cyberbullying on victims across multiple studies. A noteworthy relationship between one's participation in cyberbullying and emotional disorders were reported [8], and the research showed that higher levels of cyberbullying oppression were related to higher levels of depressing affect. In the research, it was found that 32% of cyberbullying targets experienced at least one symptom of stress[8].

Youngsters cyberbullying victims experience a lot of similar damage as the person experiencing traditional bullying. These effects can include academic performance depreciation, loos of confidence. Other effects can be more saver [5]. Recent studies by [9] show that the most mutual emotional side effect that the victims experience as a result of cyberbullying is being frightened.

A study [18] to study how cyberbullying relate to the metal health challenges on high school students. Logistic regression is used to perform a test to identify significant relationship between cyberbullying and mental health of students. The research revealed a valuable result that among ethnicity, gender, and grade, the cyberbullying victimization can be used as a predictor of negative mental health consequences. Cyberbullying victimization was reported to increase the probability of drag abuse and misery by nearly 2 times, and increase the probability of committing suicide by 3.2 times.

III. THE DATA SCOURCES FOR CYBERBULLYING RESEARCH

In this section, an overview is engaged to the data sources for cyberbullying research, where the gathering and labelling of the datasets are concerned. The most popular dataset used in cyberbullying research is the formspring.me [10], where the data was collected and labelled by the Amazon Mechanical Turk online service. The formspring.me dataset contains about 13 000 text messages.

The Fundation Barcelona Media provided five datasets to the Content Analysis for the WEB 2.0 workshop (CAW 2.0), which includes the data from Myspace social network Kongrate and Slashdot. Kongrate provides web browser based online gaming, chatrooms for real-time texting amongst game players. Slashdot is a website where individuals can read and discuss technology related topics [11]. In [12] the dataset was created by scraping the YouTube comment section posted on the videos on controversial topics because they often are rich source of cyberbullying comments.

In [13] used a dataset that consists of chat transcripts that are crawled from Myspace website. Table I summarizes the available datasets for cyberbullying research.

TABLE I: DATASETS FOR CYBERBULLYING RESEARCH

Dataset	Data source	Data amount	Contained cyberbullying types
YouTube Comments [29]	YouTube web site	not specified	Not specified
CAW2.0 [25]	Not specified	402,703 direct messages, 108,171 users, 2,211,628 sender/receiver pairs.	Not specified
Twitter data [26]	Twitter API	23,641,644 tweets @hillaryclinton followers 23,610,877 tweets @realdonaldtrump followers	Not specified
ASK.fm [27]	ASK.fm web site	10,000 questions and answers	17% cyberbullying of the dataset
Twitter dataset [28]	Twitter	700 tweets	301 cyberbullying tweets 129 swear-words
CyberCrime & Twitter Data [29]	Perverted-justice dataset, Stanford University	170,019 posts 467 million Twitter posts	Not specified

IV. CYBERBULLYING DETECTION

In the effort to model and detect cyberbullying behaviour, majority of the work were done by crawling the data from "formspring.me" and the Amazon Mechanical Turk for labelling their datasets. Machine learning techniques were commonly employed, for instance, the WEKA toolkit (Waikato Environment for Knowledge Analysis) machine learning methods were used to train and test their classification model. Table II summarizes the cyberbullying detection research work.

TABLE II. CYBERBULLYING DETECTION RESEARCH WORK

Data sets	Characteristics	Media	Classifier
Comments from posts on Facebook celebrities in Bangladesh	They applied unigram, bigram and trigram features from the comments and vectorized them using CountVectorizer and TfidfVectorizer [32]	Text	LSVM
Instagram posts[33]	Image features. Edge-direction coherence vector, colour histogram, Scale-invariant feature transform (SIFT), and face (number and ratio) features	Image	LSVM
Instagram Dataset [34] posted comments	Standard image and text features, image captions, and reputation of these advanced features	Image	CNN
Not specified	LBP features, visual word in vocabulary, K-means clustering generated codebook, combined image and text analysis [31]	Image & Text	SVM
Myspace posts [35]	Gender-specific features, detection of posts that ate posted by male and the posts that are posted by female. Minumal size of the training dataset for female harassing posts.	Text	SVM
Vine dataset [36]	Media session features, profile owner features, comment-based features, N-grams	Video	Ada-Boost
Facebook dataset [19]	Parents monitored Facebook interactions	Text	
Tweets CAW 2.0	Node-level features, contextual relationship-based features, activity-based features[25]	Text	SVM WKB
Facebook and Twitter	WEKA toolkit selected features [14]	Text	SVM Naive Bayes
Website Kaggle[15]	N-grams, bad word count, the term frequency–inverse document frequency (TFIDF) score and stemming	Text	SVM
Tweets	The number of bad words, the number of words showing negative emotion, mixture of first person, second, third person pronoun or proper noun with bad word, combination of link, bad words and pronouns, words presenting negative emotion [28]	Text	SVM
ASK.fm [27]	Ordinary natural language processing tackles, word-based information of the phrase	Text	SVM

The detection of cyberbullying was mainly focused on the text classification [14]. To create a dataset, data was scraped from Facebook and Twitter social networks, because they are the most widely used communities with rich cyberbullying behaviour happens. The scraped data were firstly cleaned and prepared. Then the WEKA toolkit was employed to select features to be used. The language and the text from the comments were extracted as the features to detect cyberbullying behaviour. Several languages were acquired, then only Arabic language was considered in identifying cyberbullying, with the data with other languages were discarded. The dataset was separated into two sets, one for training the model and the other for testing. SVM and Naive Bayes algorithms were used for training the model because of their popularity in the field of text classification. It was proved that it is possible to detect Arabic language cyberbullying although the performance of their system still needs some improvement in the performance.

Using machine learning. the offensiveness and the insults of the comments in the texts from social networks were detected [15]. The dataset was collected from the website Kaggle. 4000 comments from the dataset were used as the training set, and another 2500 comments are considered as the test set. The main objective was to create a model that can predict if the comment is an offensive to a person who is part of the conversations. Features were identified for detection, such as N-grams, bad word count, the term frequency–inverse document frequency (TFIDF) score and stemming. Common machine learning algorithms were employed, such as SVM and Logistic regression.

In the effort to detect Cyberbullying, a dataset was created and crawled[16] from the "formspring.me" website with a higher percentage of cyberbullying contained. Data mining algorithms were used to develop the model to detect cyberbullying. The main focus was mainly on the recall than on the precision. The near real-time feature of Twitter platform caters the need for a faster cyberbullying detection model [17]. However, most of the existing approaches are stand-alone, sequential, time consuming and suffers single-point of failure. In the efforts to develop a faster

cyberbullying detection, a collaborative distributed design was proposed to be used for cyberbullying detection. It conquered the above mentioned limitations of sequential approaches to identify cyberbullying tweets. Two classes of cyberbullying and non-cyberbullying tweets were detected .

Cyberbullying identifying and blocking were developed [19], with a Bully Blocker app designed for parents that monitors the Facebook interactions of their adolescents and notifies them when bullying is detected. In the efforts to detect cyber-aggress comments, literature [20] devised methods to detect cyberbullying using proposed feature collection to detect cyberbullying comments. Text data was converted to feature vectors so that they can train the machine learning algorithms. N-gram, counting and TF-IDF were used to construct feature vectors. N-grams are a grouped sequence of tokens from a given text. Count is the number of times each token appears in each text string. The TFIDF is applied to evaluate the impotence of words or tokens in a document. As a result, a Recall score of 72% was obtained[20].

On a mission to identify features that can be used to detect cyberbullying victimised, who their instant friends in their online social network can then in turn be examined for cyberbullying conduct[25]. Studied a large-scale real-world tweets dataset, categorized their features into three groups: context relationship, node-level and activity features, according to their observations their model could identify cyberbullies, victims, cyberbullies suspects and victim suspects. They used features that they grouped into three categories, node-level features, local interaction, contextual social structure features and activity features. Their work was strictly focused on using the twitter dataset from the CAW 2.0 because it includes the social network and the textual content of users that are communicating these content, the dataset contained about 900,000 tweets,402,703 which are direct messages,108,171 of users and 2,211,628 of the sender/receiver pairs.

In the effort to develop an approach for determining the quantity of hateful or kindness sentiment that twitter users from their dataset can possibly express in their tweets [26], a clustering algorithm is applied to classify racist user accounts, not only based on the text that relates to racism, a given racist score and a sentiment analysis based on the score. Twitter API is employed to collect 31,888 ungues users, then the geo-networks were collected which includes friends and followers, which leads to four dataset: Tweets posted by @hillaryclinton followers of about 23,641,644 tweets, Tweets posted by @realdonaldtrump followers of about 23,610,877 tweets, Tweets using the Hastings "#imwithher" and "#makeamericagreatagain" of about 27,808,823 and Racist dataset of about 2,720 words.

On the efforts to detect cyberbullying on Indonesian language text [28] used text mining techniques for text classification and determine a text containing harassment, they used eight (8) common rules for the extraction of the features to be applied to Indonesian language. They also used tweeter dataset that had 301 cyberbullying tweets and 399 tweets labelled as "non-cyberbullying". The count of words that are negative on the dataset was 2,053 and 129 of swear-words. The features used include, The count of bad words in the tweets, The count of words having a negative sentiment, a blend of a third person proper noun or pronoun with bad word to capture the cyberbullying, mixture of link, bad words and pronouns, the count of words revealing positive sentiment, blend of a first person pronoun, negative sentiment and second person pronoun blend , blend of second pronoun with bad words, blend of first person pronoun, words revealing negative sentiment and third person pronoun or proper noun. After they get the features their applied two machine learning algorithms specifically Support Vector Machine (SVM) and K-nearest Neighbor (KNN) and compered them based on the accuracy score. One major issues about their work was that errors occurred due to the Indonesian POS tagger they applied word Embedding in the dataset that is taken from Wikipedia and the news and the second issue they have is that words in their dataset where formal Indonesian language but in tweets personalities normally use casual words

In [27] with the intention to describe the complexity task of classifying cyberbullying case by applying ordinary natural language processing tackles, which demands investigation of all the word-based detailed intel of the phrase and an cyberspace system for automatic detect and observing of cyberbullying cases on the cyberspace. They make use of the dataset from the social media platform called ASK.fm which comprises a total count of about 10,000 Q&A. About 17% of the dataset has cyberbullying cases. Due to its upheld efficiency in classification problems they make use of the Support Vector Machine (SVM) and achieve an F1-score of 76.13% and precision of about 69.0%

V. DISCUSSION

From the literature overview, one finds that there is only a limited research done on cyberbullying detection or prevention. The majority of the focus is on detecting cyberbullying from the text data rather than images, audio and videos content. Current social media platforms have evolved from text messages to multimedia content, such as images with caption(known as memes),video live streaming and audios(known as voice note). Any of the abovementioned communication media content can be used for cyberbullying.

A possible factor for a limited research work on multimedia content cyberbullying detection could be the issue of the availability of data as the review have showed that there is only limited number of datasets available for training and testing the machine learning algorithms. The most datasets available was limited to text data, where majority of researchers had to use the limited existing publicly available data or generate their own data, some even scrawled the websites.

One important observation from the literature study as you can see on the above "Table II" is the popularity of SVM classifier. Majority of work done use this classifies in the area of text classification.

VI. CONCLUSION

In this paper, literature overview study was done on the cyberbullying and its detection using machine learning techniques. It was shown that only few research were done on cyberbullying detection and also shows that majority of the focus is on detecting cyberbullying from the text data rather than more common multimedia content (such as images, audio and videos). The availability of data limits the application of machine learning methods, due to the requirement of training and testing dataset. Furthermore,

most of the available datasets are limited to text data, and the majority of researchers had to generate their own data, some even scrawled the websites.

From the overview, the following points need to get more emphasis. (1) Establishing publicly accessible dataset for cyberbullying research, including both text and multimedia content from social media platforms; (2) Developing machine learning methods for cyberbullying detection and prevention, including language and voice recognition, image/videos content understanding and classification.

REFERENCES

[1] Mangaonkar, A., Hayrapetian, A. and Raje, R., 2015, May. Collaborative detection of cyberbullying behavior in Twitter data. In Electro/Information Technology (EIT), 2015 IEEE International Conference on (pp. 611-616). IEEE.

[2] PCMAG. (2011). Study: A Quarter of Parents Say Their Child Involved in Cyberbullying. [online] Available at: https://www.pcmag.com/article2/0,2817,2388540,00.asp [Accessed 31 Jan. 2018].

[3] Parime, S. and Suri, V., 2014, March. Cyberbullying detection and prevention: Data mining and psychological perspective. In Circuit, Power and Computing Technologies (ICCPCT), 2014 International Conference on (pp. 1541-1547). IEEE.

[4] Chavan, V.S. and Shylaja, S.S., 2015, August. Machine learning approach for detection of cyber-aggressive comments by peers on social media network. In Advances in computing, communications

[5] Ncpc.org. (2018). What is Cyberbullying? – National Crime Prevention Council. [online] Available at: https://www.ncpc.org/resources/cyberbullying/what-is-cyberbullying/ [Accessed 31 May 2018].

[6] S. Nadali, M. Murad, N. Sharef, A. Mustapha and S. Shojaee, "A Review of Cyberbullying Detection . An Overview", 13th International Cmiference on Intelligent Systems Design and Applications (ISDA), 2013.

[7] B. Haidar, M. Chamoun and F. Yamout, "Cyberbullying Detection A Survey on Multingual Techniques", European Modelling Symposium, 2016.

[8] Nixon, C. (2014). Current perspectives: the impact of cyberbullying on adolescent health. Adolescent Health, Medicine and Therapeutics, p.143

[9] Giménez Gualdo, A., Hunter, S., Durkin, K., Arnaiz, P. and Maquilón, J. (2015). The emotional impact of cyberbullying: Differences in perceptions and experiences as a function of role. Computers & Education, 82, pp.228-235.

[10] K. Reynolds, A. Kontostathis and L. Edwards, "Using Machine Learning to Detect Cyberbullying", 10th International Conference on Machine Learning and Applications, 2011.

[11] D. Yin, Z. Xue, L. Hong, B. Davison, A. Kontostathis and L. Edwards, "Detection of Harassment on Web 2.0", 2009

[12] K. Dinakar, R. Reichart and H. Lieberman, "Modeling the Detection of Textual Cyberbullying", 2011.

[13] J. Bayzick, "Detecting the Presence of Cyberbullying Using Computer Software", 2011.

[14] Haidar, B., Chamoun, M. and Serhrouchni, A. (2011). Multilingual cyberbullying detection system: Detecting cyberbullying in Arabic content - IEEE Conference Publication. [online] Ieeexplore.ieee.org. Available at: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8242005 [Accessed 31 Jan. 2018].

[15] Chavan, V.S. and Shylaja, S.S., 2015, August. Machine learning approach for detection of cyber-aggressive comments by peers on social media network. In Advances in computing, communications and informatics (ICACCI), 2015 International Conference on (pp. 2354-2358). IEEE

[16] Reynolds, K., Kontostathis, A. and Edwards, L. (2015). Using Machine Learning to Detect Cyberbullying - IEEE Conference Publication. [online] Ieeexplore.ieee.org. Available at: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6147681 [Accessed 31 Jan. 2018].

[17] Mangaonkar, A., Hayrapetian, A. and Raje, R., 2015, May. Collaborative detection of cyberbullying behavior in Twitter data. In Electro/Information Technology (EIT), 2015 IEEE International Conference on (pp. 611-616). IEEE.

[18] D. Goebert, I. Else, C. Matsu, J. Chung-Do and J. Chang, "The Impact of Cyberbullying on Substance Use and Mental Health in a Multiethnic Sample", Springer Science+Business Media, LLC 2010, 2011

[19] Y. Silva, C. Rich, J. Chon and L. Tsosie, "BullyBlocker: An App to Identify Cyberbullying in Facebook", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2018.

[20] V. Chavan and S. S S, "Machine Learning Approach for Detection of Cyber-Aggressive Comments by Peers on Social Media Network", 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2015.

[21] N. Zainudin, K. Zainal, N. Hasbullah, N. Wahab and S. Ramli, "A Review on Cyberbullying in Malaysia from Digital Forensic Perspective", 2016 International Conference on Information and Communication Technology (ICICTM), 16th - 17th May 2016, Kuala Lumpur, Malaysia, 2016.

[22] W. Romsaiyud1, K. Nakornphanom, P. Prasertsilp, P. Nurarak and P. Konglerd, "Automated Cyberbullying Detection using Clustering Appearance Patterns", 2017.

[23] M. Dadvar and A. Niamir, "Adopting MaxEnt to Identification of Bullying Incidents in Social Networks", 2016 27th International Workshop on Database and Expert Systems Applications, 2016.

[24] A. Squicciarini, S. Rajtmajer, Y. Liu and C. Griffin, "Identification and characterization of cyberbullying dynamics in an online social network", 2015.

[25] C. Chelmis, D. Zois and M. Yao, "Mining Patterns of Cyberbullying on Twitter", 2017 IEEE International Conference on Data Mining Workshops, 2017.

[26] E. Lozano, J. Cede˜no, G. Castillo, F. Layedra, H. Lasso and C. Vaca, "Requiem for Online Harassers: Identifying Racism from Political Tweets", 2017

[27] Y. Foong and M. Oussalah, "Cyberbullying System Detection and Analysis", 2017 European Intelligence and Security Informatics Conference, 2017.

[28] H. Nurrahmi and D. Nurjanah, "Indonesian Twitter Cyberbullying Detection using Text Classification and User Credibility", 2018 International Conference on Information and Communications Technology (ICOIACT), 2018.

[29] K. Dinakar, R. Reichart and H. Lieberman, "Modeling the Detection of Textual Cyberbullying", International Conference on Weblog and Social Media - Social Mobile Web Workshop, Barcelona, Spain 2011, 2011.

[30] N. Willard, "Cyberbullying and Cyberthreats", 2005 OSDFS National Conference, 2005.

[31] K. Kansara and N. Shekokar, "A Framework for Cyberbullying Detection in Social Network", International Journal of Current Engineering and Technology, 2015.

[32] S. Eshan and M. Hasan, "An application of Machine Learning to Detect Abusive Bengali Text", 2017 20th International Conference of Computer and Information Technology (ICCIT), 22-24 December, 2017, 2017.

[33] H. Li, "Image Analysis of Cyberbullying using Machine Learning Techniques", 2016.

[34] H. Zhong, H. Li, A. Squicciarini, S. Rajtmajer, C. Griffin, D. Miller and C. Caragea, "Content-Driven Detection of Cyberbullying on the Instagram Social Network", Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16), 2016.

[35] M. Dadvar, F. de Jong, R. Ordelman and D. Trieschnigg, "Improved Cyberbullying Detection Using Gender Information", 2012.

[36] R. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, S. Mishra and S. Mattson, "careful what you share in six seconds:Detecting cyberbullying instances in Vine", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2015.

[37] W. Moncur, K. Orzech and F. Neville, "Fraping, social norms and online representations of self", 2016.

Discrimination of Motor Imagery Task using Wavelet Based EEG Signal Features

Clifford Maswanganyi[1], Chunling Tu[1], Pius Owolawi[1], Shengzhi Du[2]
[1]*Department of Computer Systems Engineering*
[2]*Department of Electrical Engineering*
Tshwane University of Technology
Gauteng, Pretoria, RSA
clifford.rito@yahoo.com, tclchunling@gmail.com, owolawipa@tut.ac.za, dushengzhi@mail.com

***Abstract*—In this paper, we present a state-of-the-art motor imagery brain computer interface system (BCI) based on non-invasive approach in the form of electroencephalogram (EEG) with an objective of evaluating the performance of supervised machine-learning algorithms applied on features extracted from pre-processed EEG signals. Two categories of features were utilized namely a high dimensional feature set extracted from 22 EEG channels and a feature set extracted from two EEG channels (C3 and C4). Four signal classifiers namely KNN, Regression tree, NB and LDA are applied on wavelet-based EEG signal features for discrimination of four classes of motor imagery (MI) tasks (left hand, right hand, foot and tongue). Efficient discrimination of motor imagery tasks is significantly dependent on signal-to-noise ratio of EEG signals to enhance the performance of signal classifiers. A pre-processing technique is firstly applied on filtered EEG signals to remove contamination in the form of artifacts. Then, useful signal features are extracted from artifact free EEG signals, whereby relevant subsets with high predictive power are selected using feature selection technique. The best features subsets are fed into signal classifiers for classification purposes. A highest average classification accuracy of 73.06% and 72.95% was achieved using NB while classifying both features acquired from 2 and 22 EEG channels respectively.**

Keywords—Brain-computer interface (BCI), feature extraction, Electroencephalogram (EEG), feature selection, classification algorithms

I. INTRODUCTION

Brain computer interface (BCI) has been utilized in various applications over the past decades. It has recently emerged as one of the most significant pattern recognition systems, it is utilized to assist patients suffering from diseases such as sensory impairments or severe motor disabilities to communicate with their surrounding environment [1]. BCI is an interaction between the brain and external environment through electrical signals acquired from the brain, whereby EEG signals in the form of motor imagery task (MI) are translated into computer readable commands used to control electronics equipment such as prosthetic limbs or robotic arms and motorized wheel [2]. To acquire a more precise and accurate motor imagery task, EEG signals are subjected to a series of pre-processing techniques [3]. Furthermore, relevant signal features associated with motor imagery task are extracted from artefact free signals to enhance the performance of signal classifiers [3]. However, due to non-stationary and non-linear nature of neural signals, classification performance of signal classifiers suffers from limited motor imagery task detection and low accuracy [4]. Regardless of the resources and research efforts invested in BCI systems motor imagery tasks prediction still remains one of the most significant challenge. Alqattan et al. [5] proposed a prediction of six motor imagery task by discriminating EEG signals using linear discrimination analysis (LDA) and support vector machine (SVM), a butterworth bandpass filter was firstly applied to de-noise signals then a discrete wavelet transform was used to extract features, which were also used as inputs to the two classifiers. Both signal classifiers achieved an accuracy of up to 75%. Zhuang et al. [6] proposed a prediction of four motor imagery task by discriminating EEG signals using back propagation neural network, a bandpass filter and independent component analysis (ICA) were firstly applied to eliminate artefact and a common spatial pattern to extract signal features. A classification accuracy of up to 84% was achieved.

Tayeb et al. [7] proposed a prediction of two motor imagery task by classifying EEG signals using spiking neural network, a digital butterworth FIR filter was used to pre-process EEG signals. Principal component analysis (PCA) was used to eliminate the challenge of dimensionality from extracted features, and classification accuracy of up to 75% was achieved. In this paper, we focus on evaluating the performance of machine learning techniques by discriminating four classes of motor imagery tasks namely left hand, right hand, foot and tongue. Furthermore, we introduced several techniques for each phase of our BCI system. (1) Independent component analysis (ICA) is employed to further process filtered EEG signals. (2) Wavelet packet transform extracts features from pre-processed signals. (3) Differential evolution-based channel and feature selection technique is used to select the best subsets or relevant features. (4) The best subsets are used as input to signal classifiers to effectively discriminate or predict motor imagery task.

II. MATERIALS AND METHODS

A. EEG Signal Acquisition

An online publicly available database containing dataset IIa from BCI competition IV (2008) was utilized during implementation of each of the four phases of our BCI system. BCI competition IV dataset IIa is the most widely used multiclass motor imagery movements dataset [8]. EEG

signals were acquired from the brain using twenty-two Ag/AgCl channels. All channels were positioned on the scalp of each of the nine subjects using 10-20 electrodes positioning system, with both left and right mastoids serving as ground and reference respectively as shown in Fig. 2 [8]. Recorded EEG signals were sampled with 250Hz frequency. As part of the experiment each subject was required to perform motor imagery tasks (left hand, right hand, both feet and tongue) based on a visual cue displayed on a computer screen, the visual cue represented four arrow directions namely left, right, down and up associated with each of the four motor imagery classes [8]. A fixed cross '+' was displayed on the screen at *t*=0s to signify a start of a trial. Then an external stimulus in a form of a beeping sound after two seconds (*t*=2s). A visual cue represented by an arrow pointing to four directions was displayed for *t* = 1.25s on the screen. From *t*= 3.25s until *t* = 6s, subjects were required to perform motor imagery task [8]. All dataset acquired from nine subjects consist of six runs with one run representing 48 trials. Experimentation paradigm is illustrated in Fig. 1.

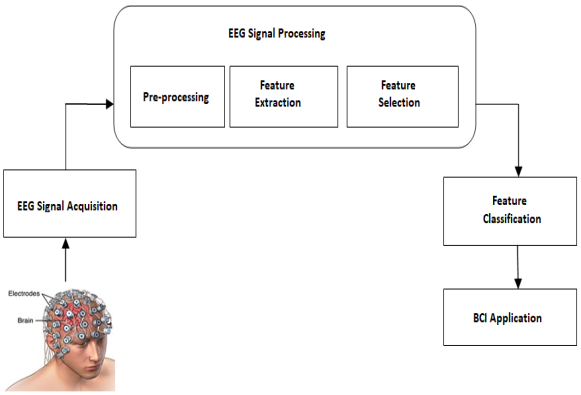


Fig. 1. Brain Computer interface block diagram

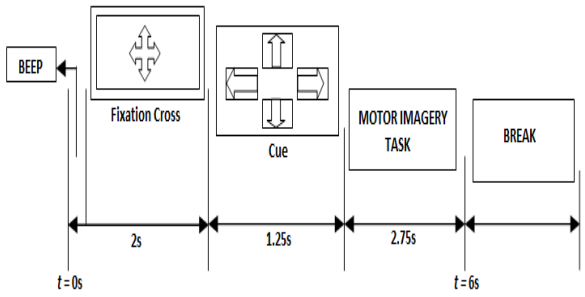


Fig. 2. Experiment Setup and Visual Stimuli

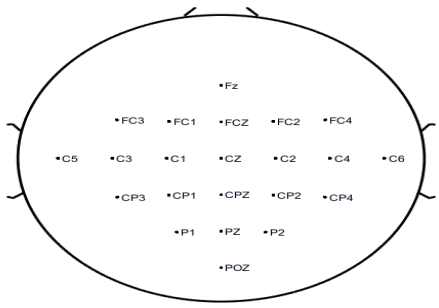


Fig. 3. 10-20 Electrodes positioning

B. EEG Signal Pre-processing

EEG signals acquired from the brain are mainly contaminated by the occurrence of physiological and non-physiological artifact during EEG recording, which may in turn reduce the signal-to-noise ratio of EEG signals resulting in a poor classification accuracy of signal classifiers [8]. Dataset IIa from BCI competition IV utilized in our BCI system was subjected to a series of pre-processing techniques, in this case a band-pass filter between 0.5Hz and 100Hz was firstly applied on EEG dataset to eradicate non-physiological artifacts [9]. Furthermore, a 50Hz notch filter was applied to eliminate the effect of line noise. A statistical technique for identifying components that are statistically separate from each other in a dataset consisting of multiple different features was applied on filtered EEG signals, in this case independent component analysis [9]. MATLAB plugin pre-processing toolbox called EEGLAB was utilized to facilitate the implementation of independent component analysis (ICA) on filtered EEG signals to eliminate the effect of physiological artifact such as EMG, ECG and EOG [9]. A mathematical ICA model is defined by (1) and (2).

$$x(t) = A.s(t) \quad (1)$$

$$y(t) = W.x(t) \quad (2)$$

Whereby $s(t)$ defined by (3) represents the original signals, with the signal length denoted by t and n representing the number of channels from which the original signals are acquired.

$$s(t) = [s_1(t), s_2(t), \dots, s_n(t)]' \in R^{n \times M}, \quad (3)$$

With each signal's sampling point denoted by *M*. With *n* channels of observed mixtures $x(t)$denoted by (4):

$$x(t) = [x_1(t), x_2(t), \dots, x_n(t)]' \in R^{n \times M}. \quad (4)$$

C. Feature Extraction

MATLAB's wavelet toolbox was utilized to facilitate the extraction of signal features from processed EEG signals. A wavelet based approach which is an extension of discrete wavelet transform (DWT) known as wavelet packet transform (WPT) was used to decompose processed EEG signals into five decomposition levels, whereby five decomposition level resulted in each signal being broken down into five detail wavelet coefficients and five approximation wavelet coefficients from which features are extracted [10].

WPT is a signal decomposition technique used to decompose signals into wavelet coefficients using a series of low-pass and high-pass filters, and it is highly efficient in displaying a more informative high and low frequencies of a signal [11]. WPT defined by (5), firstly decomposed signals into approximation coefficients $C_{N,k}$ and detail coefficients $d_{j,k}$using 4th order Daebuchies (db4) mother wavelet, with $\varphi(t)$ and $\theta(t)$ representing a wavelet function and a scaling function respectively [11]. Every decomposition level is denoted by two output signals representing detail (6) and approximation (7) coefficients, with both $h[.]$ and $l[.]$ respectively representing a high-pass filter and a low-pass filter used to filter the signal $x[k]$.

$$x(t) = \sum_{k=-\infty}^{+\infty} C_{N,k}\theta(2^{-N}t - k) + \sum_{j=1}^{N}\sum_{k=-\infty}^{+\infty} d_{j,k}2^{\frac{-j}{2}}\varphi\left(2^{-j}t - k\right) \quad (5)$$

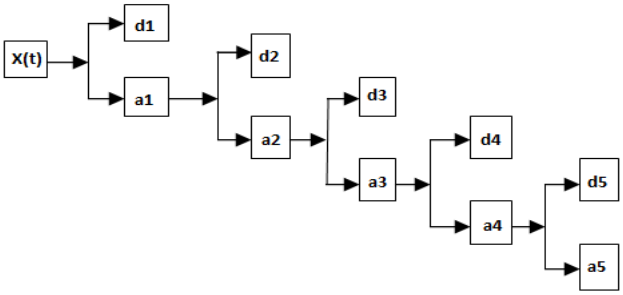


Fig. 4. 5 level signal decomposition

$$D_j[i] = \sum_k x[k].h[2.i - k] \quad (6)$$

$$A_j[i] = \sum_k x[k].l[2.i - k] \quad (7)$$

Wavelet toolbox's build-in function (mdwtdec) is used to decompose EEG signals into wavelet packet transform tree using five decomposition levels, and an overlapping window technique is then utilized to extract features from decomposed wavelet coefficients, decomposed EEG signals are separated into several windows with each window size set to 32 and each window incremented by 32 or 32 spaces apart [11]. A high dimensional feature set is firstly extracted from all 22 EEG channels, as a result 63 features are extracted from each of the 22 EEG channels resulting in a 6552x1387 feature matrix. In this case, 6552 represent the number of samples and 1387 represent the number of features. Another feature set is also extracted from only two EEG channels (C3 and C4) where most of the motor imagery tasks occur and 63 features are extracted per channel, resulting in a 6552x127 feature matrix.

D. Feature Selection

To eliminate redundant features and the challenge of dimensionality from the extracted feature vectors we applied differential evolution-based channel and feature selection technique (DEFS) [12], this technique is based on two methods namely differential evolution (DE) optimization and a repair mechanism that is centered on feature distribution mechanism [12]. DEFS is a feature selection technique utilized to identify the best subsets consisting of different dimensionality from extracted features, this technique is highly efficient in detecting subsets that are more informative and have high predictive power [12]. To identify the best subsets from both sets of extracted features, several input parameters are assigned to DEFS algorithm. Our extracted feature sets are divided into training and testing set with the last column of our features being the class labels, and then the desired number of features was set to ten. Furthermore, the population size is set to 50 and the number of iterations is set to 100 as our terminating condition.

E. Classification

In this section, four supervised machine-learning algorithms are utilized to classify four classes of motor imagery tasks, which include imaged left hand, right hand, foot and tongue. Two categories of features extracted from 22 EEG channels and 2 EEG channels respectively are used as inputs to the DEFS algorithm, which select the best subsets with high predictive power from all features. The best feature subsets acquired from DEFS algorithm are then used as inputs to four signal classifiers. K-NN which is a pattern recognition technique utilized to classify samples into classes by locating k samples containing similar independent variables in the training set, whereby KNN calculate the Euclidean distance between samples based on independent variables to identify samples that are closest in the training set [13].

NB which is a statistical technique based on Bayes theorem, it is utilized to solve classification problem by predicting the probability of a class generating observed value for features and assumes that an individual feature of a certain class is isolated from other features [13]. LDA which is the most commonly utilized pattern recognition technique to solve classification problem through identifying linear combination of features by forming a maximum isolation of objects comprising of different behaviours [14]. Regression tree is a supervised machine learning algorithm utilized to predict the probability in this case of motor imagery tasks, whereby chances of binary splits occurrence on extracted features is evaluated by identifying slits with a minimum squared error, and this process occurs continuously across the nodes until a terminating condition is satisfied [15]. The performance of each of the signal classifiers are evaluated based on the accuracy acquired through classification of the two sets of features.

III. RESULTS AND DISCUSSION

The proposed work utilizes BCI competition IV dataset IIa with MATLAB utilized as a platform to facilitate implementation of each of the algorithms applied on each and every phase of the BCI system. Four machine learning algorithms are evaluated based on classification accuracy acquired through classifying features extracted from five subjects. Table 1 shows the classification accuracy of four signal classifiers applied on features acquired from 22 EEG channels during discrimination of four motor imagery task (left hand, right hand, foot and tongue). The results show that K-NN algorithm achieved the lowest average classification accuracy of 70.88% with the highest average classification accuracy of 72.95% achieved by NB algorithm. Table 2 shows the classification accuracy of four signal classifiers acquired from only 2 EEG channels during prediction of four motor imagery tasks, with the lowest average accuracy of 71.22% achieved using K-NN algorithm and a highest average accuracy of 73.06% achieved using NB algorithm. The classification experimentation results depicted in TABLE I and TABLE II illustrate that NB algorithm outperformed the other three signal classifiers during discrimination of both feature sets. TABLE III shows that features acquired from 2 EEG channels has a slightly higher classification accuracy as compared to features acquired from 22 EEG channels.

TABLE I. CLASSIFICATION RESULTS FOR FEATURES ACQUIRED FROM 22 EEG CHANNELS

Subjects	LDA	K-NN	NB	Regression Tree
S1	70.54%	70.36%	72.70%	72.58%
S2	72.19%	71.97%	73.51%	73.39%
S3	68.48%	67.68%	72.31%	71.95%
S4	72.70%	71.72%	73.40%	72.99%
S5	72.77%	72.70%	72.83%	72.90%

TABLE II. CLASSIFICATION RESULTS FOR FEATURES ACQUIRED FROM 2 EEG CHANNELS

Subjects	LDA	K-NN	NB	Regression Tree
S1	70.57%	70.51%	72.86%	72.92%
S2	73.36%	72.56%	73.33%	73.36%
S3	67.50%	66.41%	71.54%	70.27%
S4	73.31%	72.93%	73.95%	73.88%
S5	73.40%	73.69%	73.66%	73.75%

TABLE III. AVERAGE CLASSIFICATION ACCURACY

Classifier	2 EEG channels	22 EEG channels
LDA	71.63%	71.33%
K-NN	71.22%	70.88%
NB	73.06%	72.95%
Regression Tree	72.84%	72.76%

IV. CONCLUSION

In this study, we presented an EEG based BCI system to evaluate the performance of signal classifiers applied on two sets of extracted features. Four signal classifiers are applied on both high dimensional feature set acquired from 22 EEG channels and a feature set acquired from 2 EEG channels using wavelet packet transform. To minimize the challenge of dimensionality and redundant features we applied DEFS technique on both feature sets. Furthermore, DEFS was used to select only the best subsets with high predictive power for classification of four motor imagery tasks. We evaluated the performance of signal classifiers based on the classification accuracy acquired from discrimination of motor imagery tasks. The results from our classification experiment indicate that signal classifiers applied on feature set acquired from two EEG channels yield a slightly higher accuracy than signal classifiers applied on high dimension feature set acquired from 22 EEG channels.

REFERENCES

[1] Y. U. Khan, and F. Sepulveda, "Brain–computer interface for single-trial EEG classification for wrist movement imagery using spatial filtering in the gamma band," IET Signal Process, vol. 4, pp. 510-517, 2010.

[2] A. Myrden, and T. Chau, "A passive EEG-BCI for single-trial detection of changes in mental state," IEEE Transactions on neural systems and rehabilitation Engineering, vol. 25, pp. 345-356, 2017.

[3] R. Masoomi, and A. Khabem, "Enhancing LDA-based discrimination of left- and right-hand motor imagery: outperforming the winner of BCI competition II, " In: 2015 2nd international conference on knowledge-based engineering and innovation, Tehran, Iran, 2015, pp. 392-398.

[4] B. Hu, X. Li, S. Sun, and M. Ratcliffe, "Attention Recognition in EEG-Based Affective Learning Research Using CFS+KNN Algorithm, "IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 15, pp. 38-45, 2018.

[5] H. Suk, and S. Lee, "A Novel Bayesian Framework for Discriminative Feature Extraction in Brain-Computer Interfaces, "IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 35, pp. 286-299, 2013.

[6] D. Vautrin, X. Artusi, M.O. Lucas, and D. Farina, "A Novel Criterion of Wavelet Packet Best Basis Selection for Signal Classification with Application to Brain–Computer Interfaces," IEEE Transactions on Biomedical Engineering, vol. 56, pp. 2734-2738, 2009.

[7] D. Coyle, G. Prasad, and T. M. McGinnity, "A Time-Series Prediction Approach for Feature Extraction in a Brain–Computer Interface, "IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 13, pp. 461-467, 2005.

[8] C. Vidaurre, M. Kawanabe, P. von Bunau, B. Blankertz, and K. R. Muller, "Toward Unsupervised Adaptation of LDA for Brain–Computer Interfaces," IEEE Transactions on Biomedical Engineering, vol. 58, pp. 587-597, 2011.

[9] L. F. Nicolas-Alonso, R. Corralejo, J. Gomez-Pilar, D. Álvarez, and R. Hornero, "Adaptive Stacked Generalization for Multiclass Motor Imagery-Based Brain Computer Interfaces," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 23, pp. 702-712, 2015.

[10] J. Guo, S. Gao, and B. Hong, "An Auditory Brain–Computer Interface Using Active Mental Response," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 18, pp. 230-235, 2010.

[11] B. Kamousi, Z. Liu, and B. He, "Classification of Motor Imagery Tasks for Brain-Computer Interface Applications by Means of Two Equivalent Dipoles Analysis, "IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 13, pp. 166-171, 2005.

[12] L. He, D. Hu, M. Wan, Y. Wen, K. M. von Deneen, and M. Zhou, " Common Bayesian Network for Classification of EEG-Based Multiclass Motor Imagery BCI, " IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 46, pp. 843-854, 2016.

[13] A. Sharmila, and P. Geethanjali, "DWT Based Detection of Epileptic Seizure from EEG Signals Using Naive Bayes and k-NN Classifiers," vol. 4, pp. 7716-7727, 2016.

[14] R. N. Khushaba, S. Kodagoa, S. Lal, and G. Dissanayake, "Driver Drowsiness Classification Using Fuzzy Wavelet Packet Based Feature Extraction Algorithm," IEEE Transaction on Biomedical Engineering, vol. 58, pp. 121-131, 2011.

[15] Z. Zhang and K. K. Parhi, "Seizure Detection using Regression Tree Based Feature Selection and Polynomial SVM Classification," In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Milan, Italy, 2015, pp. 6578-6581.

Gender Influences on Preference of Pedagogical Interface Agents

Ntima Mabanza
Department of Information Technology
Central University of Technology (CUT), Free State
Private Bag X20539, Bloemfontein 9300, South Africa
nmabanza@cut.ac.za

Abstract*—The purpose of this paper was to examine gender influences on the preference of Pedagogical Interface Agents (PIAs) amongst a group of male and female adult computer literacy trainees. This was a comprehensive study with participants drawn from a township community in Bloemfontein, South Africa and matched with descriptors such as little or no formal post-school education, little or limited previous experience in computers and able to understand and speak English. Therefore, this study sought to investigate whether there is an association between the gender of adult computer literacy trainees and the preferred gender of PIAs. A total of 72 adults, specifically 41 females and 31 males participated. They were trained and given tasks to perform using Simulator Word Processor System (SMOS) which incorporated 10 PIAs that differ in terms of their human-like features such as of appearance, gender, voice and reality. A post-test questionnaire was used to measure participants' personal opinions about their preferences regarding each PIA that they interacted with. Similarly, Fischer's exact test was done to determine if there was an association between the gender of the participants and the preferred gender of the PIA. The post-test questionnaire findings indicated that there were only slight differences regarding the number of participants (females and males) who preferred one type over another one. It was also noted that human-like features of PIAs could influence the manner in which participants related to a particular PIA and their preference level of a particular PIA. Additionally, a Fischer's exact test found no statistically significant effect between participants' gender (male or female) and the preferred gender of PIA. It is, therefore concluded that the gender of participants (i.e. male or female) has no role on their preferences regarding the types, features, or gender of a PIA, neither on their preference level of a PIA.

Keywords—SMOS, adults, PIAs, preference, association, levels, previous computer experiences, trainees, gender

I. INTRODUCTION

Literature reveals that there are many proposed definitions for preference. Nevertheless, the most typical definition of preference cited by many scholars is the one proposed by Scherer [1]. Hence, for that reason the definition for preference opted for regarding the purpose of this study is the one proposed by Scherer [1]. Scherer [1] defined preference as comparatively stable evaluative judgments in the sense of liking or disliking an object over other objects. Referring to the adopted definition, it can be inferred that judgments have a profound effect on influencing preferences. Preferences are ubiquitous in real life and can be unconditional or conditional [2]. Several different factors affect our preferences [3]; one of these is gender.

According to the World Health Organisation [4], gender refers to the socially constructed characteristics of women and men – such as norms, roles and relationships of and between groups of women and men. Gender is considered to be the most important basis of human differentiation [5]. Gender differences have remained ubiquitous [5, 6]. The literature on gender differences can be classified into two main groups. The first argues that men and women are very different, and the second group maintains that the difference between men and women is slight, irrelevant and they share many similarities.

PIAs are pieces of software that portray humanlike features and can speak the natural language to guide users in a socially engaging manner through the learning process in a multimedia learning environment [7]. Collins [8] identified dual purposes for the attribution of human-like features to PIAs. Firstly, is to create a kind of human-like interaction between the user and the computer; and secondly is to improve tasks automation. The human-like features of PIAs relate to appearance such as facial expression, movement and voice of the PIA, as well as PIA characteristics such as gender, age, ethnicity, reality and tone of voice [9, 10].

In this study, the multimedia learning environment is the setting in which PIAs are used to facilitate computer training (word processing training) via narrated instructions given by different kinds of PIAs while interacting with adult computer literacy trainees through on-screen text, text with voice or animated characters appearing on a computer screen. These various kinds of PIAs differ in terms of their human-like features such as appearance, gender, the tone of voice and movements. Grayson and Newton [11] pointed out that gender influences preferences across a variety of contexts. Also, Grimbeek and Bartlett [12] claimed that gender impacts on judgements in various ways. According to Byrne and Nelson [13], the similarity-attraction hypothesis stated that humans are more attracted to others who appear and behave similarly to themselves. In line with that, Human-Computer Interaction (HCI) scholars such as Moreno and Flowerday [14], as well as Reeves and Nass [15] suggested that since computer users attribute human-like features or behaviours to computers, the

similarity-attraction hypothesis may also be applicable in the computer-based learning environment.

Researches have been carried out on the use of PIAs in multimedia learning environment over the years. Below are a few examples of these previous studies. The results from several previous studies investigating PIA gender found that male agents outperformed female agents in terms of learning outcomes [16, 17, 18] and positive evaluations of the learning experience [16, 19, 18] in math and technical domains.

Previous studies on agent similarity effects about agent gender [20, 21, 22, 23, 24, 14, 19] concluded that there was no evidence of positive impact of gender matching on learning outcomes.

The results of the study conducted by Shiban, Schelhorn, Jobst, Hornlein, Puppe, Pauli and Muhlberger [25] which investigated the effect of agent appearance features showed that female agents (rated young and attractive) outperformed male agents (rated old and less attractive) in terms of students' interest in the course material, their motivation and their performance in the exam.

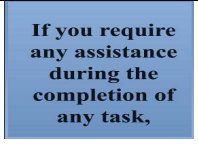
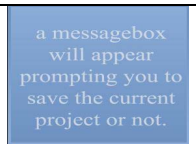
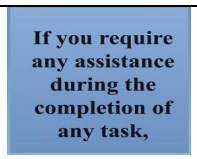
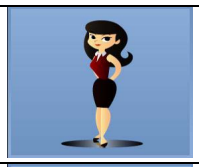
The key similarity about the past studies is that their focus was more on childhood to undergraduate student population [7]. As a result, there are few studies which investigated the use of multimedia learning environment in adult education setting, leaving a gap in the literature that needs to be filled. Hence, it can be useful to investigate these issues and fill that literature gap. In line with that, the goal of this study was to investigate the degree to which the gender of adult computer literacy trainees influences their preference of PIAs' gender.

II. METHODS

The group of adult computer literacy trainees participating in this study were drawn from the population of previously disadvantaged and unemployed adults from a township community in Bloemfontein. This study utilised two methods namely Simulator Word Processor System (SMOS) which incorporated ten (10) different kinds of PIAs (Table I) and a self-report questionnaire to obtain data related to the goal of the study. To familiarise participants with SMOS, they were all trained on how to use PIAs incorporated in SMOS. After the training, participants were given basic word processor tasks to perform using SMOS. Later, they filled in a self-report questionnaire which measured their personal opinions with regards to each of the 10 PIAs that they interacted with. The self-report questionnaire posed several questions about PIA preferences. Those preference questions consisted of several forced-choice items requiring participants to indicate their preferences for PIAs on three dimensions: PIA type preference, particular human-like feature (i.e. appearance, the tone of voice, body movements) of PIA preference, and overall PIA preference.

Data obtained using the self-report questionnaire were analysed using quantitative and qualitative data analysis techniques. Frequencies obtained for each PIA characteristic dimension preference were analysed. A comparative statistical test (i.e. Fischer's exact test) was done to determine if participants' gender has a significant influence on their preference of PIAs' gender.

TABLE I. PIAs INCORPORATED IN SMOS

PIA	Image	Characteristics
Female text (Agent1)		This agent is displayed using a solid animated white text that fades in and out, sentence by sentence.
Male text (Agent 2)		Unlike the female text agent, the male text agent is displayed using a solid animated black text that fades in and out, sentence by sentence.
Female text & audio (Agent 3)		This agent is displayed using a solid animated white text that fades in and out, while a female voice reads the text aloud, sentence by sentence.
Male text & audio (Agent 4)		This agent is displayed using a solid animated black text that fades in and out, while as a male voice reads the text aloud, sentence by sentence.
Female cartoon human (Agent 5)		This is an animated female human cartoon with eyes and mouth that are animated in correspondence with a female voice.
Male cartoon human (Agent 6)		This is an animated male human cartoon with eyes and mouth that are animated in correspondence with a male voice.
Female cartoon dog (Agent 7)		This is a female-looking cartoon dog, white and light brown in colour, with animated eyes and mouth moving in correspondence with a female voice.
Male cartoon dog (Agent 8)		This is a male-looking cartoon dog, brownish in colour, with animated eyes and mouth moving in correspondence with a male voice.
Female realistic dog (Agent 9)		This is a white realistic dog having a female voice with animated eyes and mouth moving in correspondence with an agent's voice.
Male realistic dog (Agent 10)		This is a white realistic dog having a male voice with animated eyes and mouth moving in correspondence with an agent's voice.

III. RESULTS AND DISCUSSION

The study participant group was made up of both males and females. Out of 72 study participants, 57.0% were females, and the remaining 43.0% were males. It was found that their age group varied and half of them (50.0%) belong in the age group 20-30 years. Although English was not their home language, all participants were able to speak and understand English. Regarding the qualification, 68.0% of participants were matric certificate holders, Standard 9 (22.0%), and Standard 8 (6.0%). Only 2.0% passed Grade 9, 1.0% had an N3 engineering qualification, and another 1.0% did not specify their qualification level.

To sum up, data related to Study participants revealed that the number of female participants was dominant, with most of them belonging in 20-30 age group, were able to speak and understand English, and were in possession of matric certificate as highest qualification.

SMOS incorporated different types of PIAs (see Table I). Table II presents a summary of participants' viewpoints about their preferences for each type of the different PIAs that they interacted with. Referring to Table II, *N* represents the total number of respondents and *n* is the number of respondents who responded *'Yes'* to a particular statement.

TABLE II. PREFERRED TYPE OF PIA

Statement	N	Response = *Yes*	
		n	%
I prefer a male agent to a female agent.	68	33	**49.0%**
I prefer a cartoon agent to a realistic agent.	65	33	**51.0%**
I prefer a dog agent to a human agent.	69	29	42.0%
I prefer a text agent to text and audio agent.	68	31	46.0%

The preferred type of PIA results as summarised in Table II indicate that there were only small significant differences for the number of participants who preferred one type of PIA over another one (i.e. a cartoon agent was preferred to a realistic agent, and a male agent to a female agent). It can be deduced that small variations were noted in the manner in which participants preferred individual types of PIAs incorporated in SMOS. However, no matter the PIA's levels of preference differences, all participants were still able to interact with each one of these PIAs. It shows that each type of different PIAs incorporated in SMOS was somehow usable. This implies that, to some extent PAIs fulfilled their main function which is to guide or assist users in carrying out their computing tasks which is Word processor tasks in this context. This finding is in line with the claim by Collins [9] that the attribution of human-like features to PIAs has dual purposes, which are to create a kind of human-like interaction between the user and the computer; and to improve tasks automation.

PIAs incorporated in SMOS differ in terms of their human-like features such as appearance, gender, the tone of voice and movements. Hence, every participant was asked to further select his/her most preferred human-like feature (i.e. appearance, gender, the tone of voice or movements) of every individual PIA that he/she interacted with. The summarised results of PIAs' most preferred human-like features are shown in Table III. *N* in Table III refers to the number of respondents and *n* is the number of respondents who selected a particular human-like feature of PIA (i.e. appearance, the tone of voice or movement).

TABLE III. PREFERRED FEATURES OF PIAS

PIA	Attribute					
	Appearance (N = 56)		Voice (N = 62)		Movement (N = 44)	
	n	%	n	%	n	%
Female text agent (Agent 1)	**11**	**20.0**	7	11.3	2	5.0
Male text agent (Agent 2)	5	9.0	9	15.0	2	5.0
Female text & audio (Agent 3)	6	11.0	**11**	**18.0**	3	7.0
Male text & audio (Agent 4)	5	9.0	4	7.0	3	7.0
Female cartoon human (Agent5)	5	9.0	5	8.1	**10**	**23.0**
Male cartoon human (Agent 6)	5	9.0	7	11.3	4	9.1
Female cartoon dog (Agent 7)	4	7.1	5	8.1	5	11.3
Male cartoon dog (Agent 8)	9	**16.1**	**10**	**16.1**	**7**	**16.0**
Female realistic dog (Agent 9)	4	7.1	2	3.2	2	5.0
Male realistic dog (Agent 10)	2	4.0	2	3.2	6	14.0

The results displayed in Table III suggest the following as per participants' preferred human-like features (i.e. appearance, gender, the tone of voice and body movements) of each PIA that they interacted with:

- Regarding appearance, female text (Agent 1) and male cartoon dog (Agent 8) were the two most preferred.
- With regards to the tone of voice, both female text & audio (Agent 3) and male cartoon dog (Agent 8) were most preferred.
- For movement, female cartoon human (Agent 5) and male cartoon dog (Agent 8) were the two most preferred.

Referring to the findings listed above, it can be assumed that human-like features of PIAs could have a major influence on how participants (i.e. male or female) related to a particular PIA, also can influence participants' preference level of a particular PIA. This implies that participants can create a certain kind of personal subjective impressions (i.e. positive or negative) based on human-like features of PIAs. Furthermore, participants' subjective impressions (judgment) can, to some extent, can differ from one kind of PIA to another. It is

suggested that in order PIAs to better fulfil their tasks (i.e. guide or assist users through the learning process in a multimedia learning environment), consideration must be given to the different kinds of human-like features attributed to a particular PIA.

Every participant was also asked to indicate his/her first choice of PIA regarding overall preference. The overall preference consisted of choosing a particular PIA based on a combination of all three human-like features (i.e. appearance, gender, the tone of voice and movements) that it portrayed. Table IV provides a summary of the overall preference results of the 10 PIAs that participants interacted with. Referring to the results displayed in Table IV, note that *N* indicates the number of respondents and *n* is the number of respondents who selected a particular PIA across all three human-like features (appearance, the tone of voice and body movements).

TABLE IV. OVERALL PREFERENCE OF PIAS

PIAs	Overall Preference (N = 69)	
	n	%
Female text agent (Agent 1)	9	**13.1**
Male text agent (Agent 2)	6	9.0
Female text & audio (Agent 3)	8	**12.0**
Male text & audio (Agent 4)	5	7.3
Female cartoon human (Agent 5)	7	10.1
Male cartoon human (Agent 6)	7	10.1
Female cartoon dog (Agent 7)	6	9.0
Male cartoon dog (Agent 8)	15	**22.0**
Female realistic dog (Agent 9)	2	3.0
Male realistic dog (Agent 10)	4	6.0

Data in Table IV reveal that the overall preferred PIAs as selected by participants across all three human-like features portrayed by PIAs (i.e. appearance, voice and movement) were the male cartoon dog (Agent 8), the female text (Agent 1), and the female text & audio (Agent 3). The outcome is consistent with previous work [16, 17, 18] who found that male agents outperformed female agents regarding learning outcomes. Also, the outcome is also in line with the findings [16, 19, 18] that male agents outperformed female agents regarding positive evaluations of the learning experience in technical domains.

The study participant group was made up of both males and females. However, every participant either male or female was able to select his/her preferred type of PIA, most preferred feature of PIA, and the overall preference of PIAs (see respectively Table II, Table III, and Table IV). Hence, it was necessary to investigate further if the mere fact that of being a male or a female could have an association with the preferred gender of PIA. In this respect, a Fisher's exact test was performed to determine if there was an association between the gender of participants and the preferred gender of the agent. The results of a Fisher's exact test are displayed in Table V. Regarding the results in Table V, *N* signifies the number of respondents as per their gender and *n* is the number of respondents who selected male PIA.

TABLE V. PARTICIPANTS' GENDER AND PREFERRED GENDER OF PIA

Statistic	Participants' Gender	
	Male (N=28)	Female (N=40)
Participants who prefer male PIA		
n	16	17
%	57.0	43.0
P-value	**0.3246**	

The results in Table V indicate that as per participants' gender 57.0% of male preferred a male agent, whereas only about 43.0% of female preferred a male agent. The Fisher's exact result (P = 0.3246) as per Table V reveals that this difference is not statistically significant. It is inferred that no significant difference was found between the proportion of male participants who preferred a male agent to a female agent and the proportion of female participants who preferred a male agent to a female agent. This implies that participants' gender (i.e. male or female) has no significant effect on their preference for PIAs' gender. This finding is in line with [20, 21, 22, 23, 24, 14, 19] who also found there was no evidence of positive impact of gender matching on learning outcomes. It is therefore argued that males and females have the likelihoods of displaying the same kinds of reactions, attitudes (i.e. positive or negative) with regards to humanlike features of a particular PIA.

IV. CONCLUSION

This study sought to investigate whether there is an association between the of adult computer literacy trainees' gender and preferred gender of PIAs. There was a total of 41 women (57.0%) and 31 men (43.0%) from a township community in Bloemfontein. This group of participants was trained and assessed using SMOS incorporating ten (10) types of PIAs (Table II) which differ regarding their human-like features such as appearance, gender, the tone of voice and movements. A post-test questionnaire was used to measure participants' personal preferences with regards to each of the individual PIA that they interacted with. The findings of this study indicate that, in general, there were only slight differences regarding the number of participants (females and males) who preferred one type over another one (see Table II). In this study, it was noted that PIAs' human-like features could influence the manner in which participants (i.e. males or females) related to a particular PIA and participants' preference level of a particular PIA. The result of the Fischer's exact test conducted in this study (Table V) found no statistically significant effect between participants' gender (male or female) and their preferred gender of PIAs. It is concluded that the gender of participants (i.e. male or female) has no role (impact)

on their preferences regarding the type, features, or gender of PIA, neither on their preference level of PIA.

ACKNOWLEDGMENT

The author would like to thank the adult computer literacy trainees for their willingness to participate in this study and also the SMOS developer, Mr Louwrens Potgieter. Many thanks go to the Telkom Centre of Excellence (CoE) at the Department of Computer Science and Informatics, University of the Free State for partially funding this study. Lastly but not least, a hearty thanks to Mr Casper Wessels for supplying the computers that were used to carry out this study.

REFERENCES

[1] K.R. Scherer, "What are emotions? And how can they be measured?" *Social Science Information*, Vol. 44, pp. 695-729, 2005.

[2] F. Rossi, "Preference reasoning", In: Logic Programming, Lecture Notes in Computer Science, vol 3668, M. Gabbrielli and G. Gupta, Eds. Berlin: Springer, 2005.

[3] M. Vabo and H. Hansen, "The Relationship between food preferences and food choice: a theoretical discussion", *International Journal of Business and Social Science*, Vol. 5, No. 7, pp. 145-157, 2014.

[4] World Health Organization, "Gender", *World Health Organization* 2018. [Online]. Available at: http://www.who.int/gender-equity-rights/understanding/gender-definition/en/ [Accessed: September 22, 2018].

[5] M. Harper and W.J. Schoeman, "Influences of gender as a basic-level category in person perception on the gender belief system", *Sex Roles*, Vol. 49, Issue 9–10, pp. 517-526, 2003.

[6] T. Buser, M. Niederle and H. Oosterbeek. "Gender, competitiveness, and career choices", *The Quarterly Journal of Economics*, Vol. 129, no. 3, pp. 1409-1447, 2014.

[7] N. Mabanza and L. De Wet, "Determining the usability effect of pedagogical interface agents on adult computer literacy training", In: E-Learning Paradigms and Applications, Studies in Computational Intelligence, Vol. 528, M. Ivanovic´ and L.C. Jain, Eds. Berlin: Springer-Verlag, 2014, pp. 145-182.

[8] R. Collins, "Running head: anthropomorphic design", *R. Collins* 2007. [Online]. Available at: http://www.ruthcsmith.com/GWU%20Papers/AnthropomorphicAgents.pdf [Accessed: September 22, 2018].

[9] G. Ozogul, A. M. Johnson, R. K. Atkinson and M. Reisslein, "Investigating the impact of pedagogical agent gender matching and learner choice on learning outcomes and perceptions", *Computers and Education*, vol. 67, pp. 36-50, 2013.

[10] L.T. Tien and K. Osman, "pedagogical agents in interactive multimedia modules: issues of variability", *Procedia Social and Behavioral Sciences*, vol. 7, pp. 605-612, 2010.

[11] T. S. Behrend, L. F. Thompson, A. W. Meade, M. S. Grayson and D. A. Newton,"Gender differences in career choice influences". Presented at the 22nd Annual Meeting of the Society for Industrial and Organizational Psychology, April 2007, New York, USA. 2007.

[12] Grimbeek, P. y Bartlett, B. J. (2005). "The effect of gender on judgments about the future prospects of a baby girl or boy: Does our gendered culture also influence judgments about the future?", Stimulating the 'Action' as Participants in Participatory Research, Vol. 2, B. Bartlett, F. Bryer and D. Roebuck, Eds. Nathan, Qld.: Griffith University, 2005, pp. 94-100.

[13] D. Byrne and D. Nelson, "Attraction as a linear function of proportion of positive reinforcements". *Journal of Personality and Social Psychology Bulletin*, vol. 4, pp.240-243, 1965.

[14] R. Moreno and T. Flowerday, "Students' choice of animated pedagogical agents in science learning: a test of the similarity attraction hypothesis on gender and ethnicity", *Contemporary Educational Psychology*, vol. 31, pp. 186-207, 2006.

[15] B. Reeves and C. Nass, *The media equation: how people treat computers, television, and new media like real people and places*, Cambridge, MA: Cambridge University Press, 1996.

[16] I. Arroyo, B. P. Woolf, J. M. Royer and M. Tai, "Affective gendered learning companion", In International Conference on Artificial Intelligence and Education, 2009, Brighton, England: IOS Press, 2009.

[17] Y. Kim and Q. Wei, "The impact of learner attributes and learner choice in an agent-basedeEnvironment", *Computers & Education*, vol. 56, pp. 505-514, 2011.

[18] K. N. Moreno, B. Klettke, K. Nibbaragandla and A. C. Graesser, "Perceived characteristics and pedagogical efficacy of animated conversational agents", *Lecture Notes in Computer Science: Intelligent Tutoring Systems*, vol. 2363, pp. 963-971, 2002.

[19] A. L. Baylor and Y. Kim, "The role of gender and ethnicity in pedagogical agent perception", Proc. of world conference on e-learning in corporate, government, healthcare, and higher education, 2003, G. Richards, Eds. Chesapeake, VA: AACE, 2003. pp. 1503-1506.

[20] T. S. Behrend and L. F. Thompson, "Similarity effects in online training: effects with computerized trainer agents", *Computers in Human Behavior*, vol. 27, pp. 1201-1206, 2011.

[21] K. M. Lee, K. Liao and S. Ryu, "Children's responses to computer-synthesized speech in educational media: gender consistency and gender similarity effects", *Human Communication Research*, vol. 33, pp. 310-329, 2007.

[22] E. A. Plant, A. L. Baylor, C. E. Doerr and R. B. Rosenberg-Kima, "Changing middle-school students' attitudes and performance regarding engineering with computer-based social models", *Computers & Education*, vol. 53, pp. 209-215, 2009.

[23] R. B. Rosenberg-Kima, A. Plant, C. E. Doerr and A. L. Baylor, "The influence of computer-based model's race and gender on female students' attitudes and beliefs towards engineering", *Journal of Engineering Education*, vol. 99, no. 1, pp. 35-44, 2010.

[24] H. Van der Meij, J. Van der Meij, and R. Harmsen, "Animated pedagogical agents: do they advance student motivation and learning in an inquiry learning environment?", Technical Report TR-CTIT-12-02 Enschede: Centre for Telematics and Information Technology University of Twente, 2012.

[25] Y. Shiban, I. Schelhorn, V. Jobst, A. Hörnlein, F. Puppe, P. Pauli and A. Mühlberger, "The appearance effect: influences of virtual agent features on performance and motivation", *Computers in Human Behavior*, vol. 49, pp. 5-11, 2015.

Preprocessed Faster RCNN for Vehicle Detection

Mduduzi Manana, Chunling Tu and Pius Adewale Owolawi
Department of Computer Systems Engineering
Tshwane University of Technology
Pretoria, South Africa
mdwens@gmail.com, tclchunling@gmail.com, owolawipa@tut.ac.za

***Abstract*—This paper presents a pre-processed faster region convolution neural network (faster RCNN) for the purpose of on-road vehicle detection. The system introduces a preprocessing pipeline on faster RCNN. The preprocessing method is for the improvement on training and detection speed of Faster RCNN. A preprocessing lane detection pipeline based on the Sobel edge operator and Hough Transform is used to detect lanes. A Rectangular region is then extracted from lane coordinates which is a reduced region of interest (ROI). Results show that the proposed method improves the training speed of faster RCNN when compared to faster RCNN without preprocessing.**

Keywords—Convolution Neural Network, RPN, ROI, Hough Transform, Sobel edge detection

I. INTRODUCTION

Object detection methods have been used in many applications including that of public safety [1]. Vehicle detection methods which are based on visual object detection can help to increase driving security and decrease road crime such as on-road hijacking [2]. In this context a monocular camera is used as a passive sensor for data input and visual object detection methods are applied to detect vehicles from the traffic scene. In recent times, most state-of-the-art visual object detection methods are based on convolution neural networks (CNN) .

Region convolution neural network (RCNN) has achieved impressive detection results compared to other methods of the same class. The RCNN [3] operates by running region proposals, generated by proposal methods, on a CNN network. The drawback of RCNN is its high computational cost because each region is processed separately. Fast region convolution neural network (fast RCNN) [4] improves on the downside of RCNN by using region proposals as attention detectors for a shared feature map. The shared feature map eliminates the need of processing each region separately, this improves the computation cost of RCNN. The proposal generating method is a drawback for fast RCNN, which is slow compared to the CNN network. As a result, it has a negative impact on the overall speed of fast RCNN.

To overcome this drawback, faster RCNN [5] shares CNN computation between the proposals and the detection network. Faster RCNN creates an almost cost free region proposal generating mechanism. The Region Proposal Network (RPN) of faster RCNN generates proposals that are used as attention directors for a shared feature map. The RPN replaces an external region proposal generating process such as selective search , and this results in improved computational cost.

Faster RCNN performs well for general object detection but performs unimpressively when applied to vehicle detection. This can be improved through parameter tuning and algorithmic modification [6] .

In this paper we focus on the algorithmic modification of Faster RCNN. A preprocessing method is integrated to faster RCNN to improve training and detection speed.

The preprocessing pipeline reduces the region of interest which results in a reduced number of pixels to be processes by faster RCNN. The preprocessing pipeline is based on the Sobel edge detection [7] and Hough transform [8] with an extracted rectangular region as the end result. This improves the processing speed of the Faster RCNN. Since we are focusing on vehicles around the ego vehicle and their behavior, the tracking procedure in lane detection is ignored. This provides a simplified preprocessing method for the initial stage with computation advantages.

The preprocessed faster RCNN can then be used for vehicle detection through a vehicle mounted monocular camera which captures the traffic scene. This is for the purpose of detecting suspicious vehicles around the ego vehicle.

The rest of this paper is arranged in the following manner: Section II introduced related work, Section III presents the proposed approach, Section IV validates the proposed method by using experiments, Section V concludes the paper.

II. RELATED WORK

Vehicle detection using a camera mounted on a vehicle has been implemented in [1][9][10][11] and object detection methods were applied to detect vehicles. The improvement of a CNN based algorithm was done in [1] for vehicle detection with image pre-processing integrated. This was shown to improve computational cost by reducing the region of interest (ROI). The reduction of the ROI which is based on lane detection and tracking has been shown to be reliable on roads [10].

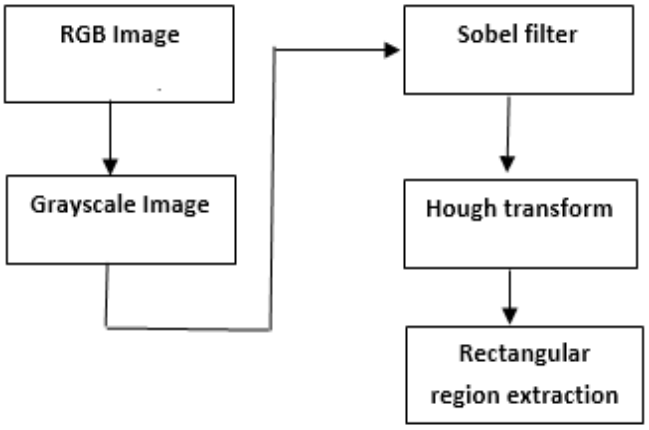


Fig. 1 Pipeline for the lane detection and rectangular region extraction

A. Lane detection

Lane detection has been used in many areas, such as lane departure warning systems. In [12] a camera based lane detection algorithm was developed to detect lane markers on the road.

Lane detection and tracking consists of the following steps: (1) detecting lane markings on the road with respect to the ego vehicle; (2) fitting those lane markings onto a model to detect the trajectory of the markings.

Edged detection methods are used to detect road markings based on pixel changes in an image. The gradient based edge detection is commonly used, where an edge is detected if the gradient is above a certain threshold. Sobel, Roberts cross operator, Prewitt operator and Canny edge [13] all comes from the gradient edge detection family. The Laplacian based edge detection is based on locating zeros at the second derivative of an image, which detects regions of rapid intensity changes indicating edges. Laplacian of Gaussian [13] is a representative of this branch. Lane marking localization also include steerable filters [14] which rotates to different angles and then synthesize those filters of arbitrary orientation from linear combinations of basis filters [10] for lane tracking and adaptive thresholds [15]. This method takes a greyscale image or color image as input then outputs a binary image representing the segmentation. For each pixel in the image a threshold is calculated.

Hough transform is used to detect straight lines in an image by connecting points that lie in a straight line. It has been used after edge detection to detect lane markers on the road [1], then other a tracking based method such as curve fitting or a lane position tracker can used as a final stage [10].

B. Vehicle detection

Object detection methods have been used in [1][9][10] for vehicle detection with region of interest reduction before vehicle detection and tracking. Conventional object detection methods were based on sliding windows to produce features that determine whether they contain objects or not. Deformable part models (DMP) [16] achieved the best results among conventional object detectors. Many vehicle detection systems get improved from DPM.

Deep learning techniques such as CNN dominate the state-of-the-art object detection methods, of which RCNN has achieved impressive detection results [3]. RCNN computes 2000 bottom-up proposals from an input image, the proposals features are extracted using CNN and each region is then classified using sector vector machine (SVM) [3]. RCNN uses selective search for proposal generation, then from each region proposal a 4096-feature vector is produced. The image is then warped into 227 x 227 pixels for computation of features. The warped image is then processed through five convolution neural networks and two fully connected layers [17]. A trained sector vector machine is then used to score the computed features.

The Fast RCNN method computes a convolutional feature map for the entire input image and then classifies each object proposal using a feature vector extracted from the shared feature map. This improves the processing speed [4]. The bottle neck of Fast RCNN is the proposal generating method which is slow compared to the detection network.

Faster RCNN addresses this bottleneck by introducing a Region Proposal Network (RPN) that shares full-image convolutional features with the detection network, and this results in nearly cost-free region proposals. RPN serves as an attention director telling the unified network where to look. A RPN is a fully convolutional network that simultaneously predicts object bounds and objectness scores at each position. The RPN is trained end-to-end to generate high quality region proposals, which are used by Fast R-CNN for detection. RPN and Fast R-CNN are then merged into a single network by sharing their convolutional features. Faster RCNN has been implemented in [6] for vehicle detection. In [6] faster RCNN is examined specifically for application in vehicle detection.

III. PROPOSED APPROACH

In this work an algorithm based on lane detection, a rectangular region of interest and faster RCNN object detection method is introduced for the purpose of vehicle detection. It begins with a lane detection pipeline for vehicles that are on the road because the lane markings are the most prevalent and visible features on the road. The lane markings are used as an estimate of the road because this is where vehicles are mostly found. This estimate enables the extraction of an accurate rectangular region of interest. The pipeline reduces the region of interest as a pre-processing step to improve the computation cost of faster RCNN. This pre-processing pipeline is similar to [1] [10] but our pipeline is quicker because there is no lane tracking involved. This provides our pipeline with a computational advantage. Method [1] also uses deep multibox which is slower compared to faster RCNN. Our lane detection pipeline reduces the pixels to be processed by faster RCNN and therefore improves on the computation cost of faster RCNN. A rectangular region of interest that is based on lane coordinates is then used as input to the Faster RCNN.

A. Lane detection

As shown in Fig. 1, the lane detection is composed of the following parts:

1) The RGB road image is converted into grayscale image.
2) The Sobel filter extracts edges.
3) The Hough transform detects straight lines.
4) Finally, a rectangular region of interest is cropped using the lines obtained from Hough transform.

The proposed pipeline uses the Sobel operator to detect edges which slightly compares to the pipeline of [1] that uses a canny edge detector which is computationally expensive. The Sobel edge detector is a simple filter with low computational cost and is ideal for the lane detection method proposed in this paper. It detects edges in a grayscale image which are then used as candidates of lines for the subsequent stage of Hough transform. The Sobel operator is considered in this situation because there is no in-depth edge detection required. Figure 2 shows the Sobel operator on an image.

The Sobel operator use two 3x3 kernels one rotated 90 degrees of the other, as shown in (1) and (2).

(a) (b)

Fig. 2 (a) greyscale image and (b) detected edges

$$Gx = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}, \quad (1)$$

$$Gy = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \quad (2)$$

The Sobel operator can be broken down as a product of an averaging kernel and a differentiating kernel (combining gradient with smoothing). *Gx* and *Gy* can be broken down into

$$Gx = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 \end{bmatrix}, \quad (3)$$

$$Gy = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 \end{bmatrix} \quad (4)$$

At each point in the image the resulting gradient can be combined to give the magnitude in (5) and direction of the edge (6)

$$G = \sqrt{Gx^2 + Gy^2} \quad (5)$$

$$\theta = \text{atan}(Gy / Gx) \quad (6)$$

This reduces the computational cost as there are fewer calculations. The direction (θ) can be set to 0 for a vertical edge detection only. This filter only needs 8 image points around a point to compute a result and then integer arithmetic is used for gradient computation. Areas of high gradient are represented as white lines on a black background which are the edges shown in Figure 2.

When the edge detection is complete, the Hough transform is applied to the detected edges to produce candidates of straight lines to represent lane markings on the road. The Hough transform is robust for eliminating noisy data on an edge image and produce accurate lines. It has an added advantage of handling incomplete data very well as a result of occlusion and shadows on the road [8]. Method on [1] uses Hough transform which it is applied to a canny edge detector.

Eq. (7) is used as a description of a line for Hough transform.

$$r = x\cos\theta + y\sin\theta \quad (7)$$

In (7) (*x, y*) represents a point (pixel) which has to be computed as to whether it belongs to a line or not. *r* represents the distance from the line to the origin. The parameter θ represents the angle between the origin of the line and the *x* axis. Each point in this context is represented by (r, θ). Every straight line that passes a single point corresponds to a sinusoidal curve in the (r, θ) plane which is unique to that point. Two or more points are considered a straight line if sinusoids are produced that intersect at (r, θ) which is a candidate of a straight line for those points. In this paper Hough transform is applied to the edging image obtained from the Sobel edge detector.

To detect a lane, a minimum length of edges is set. The coordinates of lines representing lane markings are then used to crop a rectangular region of interest as shown in Figure 3 (which shows the regular image and the cropped image). This rectangular region shown in Figure 3(b) is used as input of faster RCNN.

B. Vehicle detection

The extracted rectangular region is fed into Faster RCNN for vehicle detection. The proposed algorithm as shown in Figure 4 shows a preprocessed image being input into the convolution layer of Faster RCNN.

The purpose of the pre-processing step in Figure 4 serves to reduce the region of interest in order to speed up processing of image pixels. This will reduce training time of the detector.

The advantage of Faster RCNN is the proposal generation that is almost cost free. The region generates 300 proposals. RPN takes an image and produces rectangular proposals with an objectness score. This is done by sliding a small network over the shared feature map. At each sliding window region proposals are predicted. The regression layer outputs the coordinates of the predicted proposals and the classification layer produces a score as to the possibility of there being an object for each predicted proposal. A single-scale detection is used in this paper, because it has been shown that, for the similar performance, multi-scale detection is more expensive computationally than single scale detection [6].

(a) (b)

Fig. 3 (a) RGB image and (b) the cropped rectangular image based on points of a line detect by Hough transform

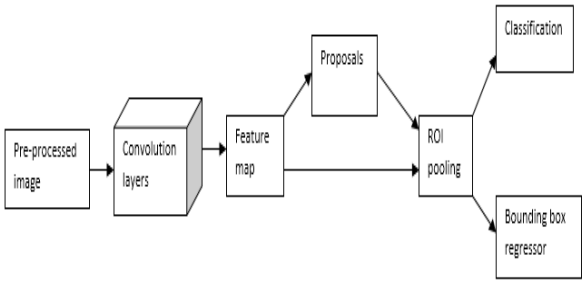


Fig. 4 Combined method with preprocessed image fig 3b based on fig. 1 and faster RCNN [5]

IV. EXPERIMENT

A similar training process for both faster RCNN and the proposed method are used to compare their training speed. Same computation resources and data are used for training. The training process is done on 102 images of vehicles on both methods. The training follows the process laid out in faster RCNN pipeline, with the following strategies:

- The regional proposal (RPN) network is first trained
- The RPN is then used to train fast RCNN
- RPN is retrained sharing weights with Fast RCNN
- Fast RCNN is retrained using the updated RPN

Training is done on a single Intel Celeron CPU, with 2GB of memory. The learning rate of the detector is 0.000001. The same data is used for the proposed method and for faster RCNN. The results of the above procedure are compared between faster RCNN and the proposed method. The results of training time, number of iterations, and mini batch accuracy are shown in Table I for the first round training, and Table II for the re-training process. The results of the final training shows that our method trains quicker than Faster RCNN with a slightly lower mini batch accuracy and more iterations.

V. CONCLUSION

A vehicle detection method is proposed that reduces a region of interest by preprocessing an image before it is processed by faster RCNN. The results show that preprocessing an image into a reduced rectangular portion of the original image improves training speed. The image is processed faster as a result of the rectangular cropped image based on lane detection. This is because with the cropped region there are a fewer pixels to process and this results in a shorter training time.

TABLE I. COMPARISON OF THE FIRST ROUND TRAINING FOR EXISTING METHODS AND PROPOSED METHOD

Method	***Iteration***	***Time elapsed(s)***	***Mini batch accuracy***
Faster RCNN (Training RPN)	72	218.13	50%
Proposed (training RPN)	67	93.33	100%
Faster RCNN (fast RCNN training using RPN)	45	81.20	81.25%
Proposed (fast RCNN training using RPN)	49	40.03	56.25%

TABLE II. RETRAINING RESULTS EXISING METHODS AND OUR PROPOSED METHOD

Method	***Iteration***	***Time elapsed (s)***	***Mini batch accuracy***
Faster RCNN (retrain RPN)	72	193.01	100%
Proposed (retrain RPN)	67	88.23	100%
Faster RCNN (fast RCNN training using updated RPN)	45	60.59	71.43%
Proposed	49	36.31	69.8%

ACKNOWLEDGMENT

This work is based on the research supported by the National Research Foundation of South Africa (Grant Number 99383).

REFERENCES

[1] H. Guan, W. Xingang, W. Wenqi, Z. Hanand W. Yuanyuan, "Real-Time Lane-Vehicle Detection and Tracking System", Beijing: Beijing Information Science & Technology University, 2016.

[2] "Crime statistics", 2017. [Online]. Available: https://africacheck.org/factsheets/factsheet-south-africas-201516-crime-statistics . [Accessed: 2 June 2017].

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Richfeature hierarchies for accurate object detection andsemantic segmentation," in Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, 2014, pp,

[4] R. Girshick, "Fast R-CNN," arXiv preprint arXiv:1504.08083, 2015.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," arXiv preprint arXiv:1506.01497, 2015.

[6] Q. Fan, L. Brown, J. Smith, "A closer look at vehicle detection for Faster RCNN", 2016 IEEE Intelligent Vehicles Symposium (IV) Gothenburg, Sweden, June 19-22, 2016.

[7] M. Wen, C. Zhong, "Application of Sobel algorithm in edge detection of images", China High-tech Enterprise, pp.57-62.

[8] J. Illingworth, J. Kittler, "A survey of Hough Transform", April 1987.

[9] Q. Shiru. L. Xu, "Research on multi-feature front vehicle detection algorithm based on video image", 28th Chinese Control and Decision Conference (CCDC), 2016.

[10] S. Sivaraman, M. Manubhai, "Integrated lane and vehicle detection localization and tracking: a synergistic approach", IEEE Transactions on Intelligent Transportation Systems, Vol. 14, No. 2, June 2013.

[11] C. L. Tu and S. Z. Du, "Moving vehicle detection in dynamic traffic contexts", in Electronics, Communications and Networks V, Singapore: Springer, 2016, pp 263-269.

[12] C. Kreucher, S. Lakshmanan, "Lana: A lane extraction algorithm that uses frequency. Domain features", IEEE Transactions on Robotics and Automation, Vol. 15, No. 2, April 1999.

[13] R. Maini, H. Aggarwal. "Study and comparison of various edge detection techniques".

[14] T. Freedman, E. Adelson, "Design and user of steerable filters", IEEE Transaction Patern Analysis and machine intelligence, vol. 13, no. 9, September 1991.

[15] M. Meuter, S. Muller-Schneiders, A. Mika, S. Hold, C. Nunn, andA. Kummert, "A novel approach to lane detection and tracking," in Proc. 12th Int. IEEE ITSC, Oct. 2009, pp. 1–6.

[16] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively Machine Intelligence", IEEE Transactions on, vol. 32, pp. 1627-1645, 2010.

[17] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012.

Composing Scalable Solver for Simulation of Electronic Circuits in SPICE

David Černý and Josef Dobeš
Department of Radioelectronics
Czech Technical University in Prague, Faculty of Electrical Engineering
Prague, Czech Republic
Emails: {cernyd1, dobes}@fel.cvut.cz

***Abstract*—The most efficient solvers use composite procedures that adaptively rearrange computation algorithms to maximize simulation performance. A similar concept can be integrated into a process of electronic circuit analysis, where the combination of different algorithms allows scalability of simulation performance. In this paper, we propose a new adaptive internal solver based on Biconjugate gradient stabilized method for the iterative solution of nonsymmetric linear systems supplemented with incomplete LU factorization as an efficient replacement for the direct solver implemented in program Spice for solving large-scale circuits. We describe basic concepts of a simulation of electronic circuits with nonlinear time dependent devices and present implementation examples of the proposed methods. Optimal setting of the method and its application in program Spice is shown in comparison to other modern iterative solvers for nonsymmetric linear systems.**

***Index Terms*—circuit simulation, SPICE, large-scale circuits, sparse matrix, BICGStab, Krylov methods**

I. Introduction

This article primarily focuses on an improvement of performance and scalability of the Spice computation core. There can be found various works suggesting modifications of algorithms with the objective to improve simulation performance or accuracy. They propose hardware specific improvements but also changes in simulation procedures such as matrix reordering techniques, in LU factorization and in numerical integration. From recent publications in this area, we highly recommend article [1], that proposes a parallel sparse matrix solver run on FPGA. The published method clearly outperforms standard direct solvers. The given material also presents a comparison of the most powerful matrix solvers, such as UMFPACK, KLU, and Kundert sparse matrix packages. Paper [2] gives another attempt to enhance simulation performance using a single GPU-accelerated LU factorization. The proposed LU factorization approach shows very impressive speed up on NVIDIA graphic card in comparison to high-performance package for solving large sparse linear systems PARADISO. Another interesting article [3] presents a reimplementation of Spice simulator on a parallelized single field-programmable gate array (FPGA). Their technique allows not only to accelerate the evaluation of analysis, but also to reduce the energy needed for computation in comparison to conventional processors.

With regards to the topic of this paper, we must cite [4]. In the paper, a replacement of the standard direct method with an iterative Krylov-subspace solver with preconditioner is proposed. As it will be revealed later, a solver of sparse matrix systems based on the iterative method can compete with a direct solver only in special situations (large, positive-definite or banded systems). Hence, it is questionable whether full replacement of the standard direct method is realistic. From the result presented in the article, it is not clear which of the proposed Krylov methods is better for simulation of electronic circuits. Instead, the article provides an unfair comparison between GMRES, QMR, BICGStab methods (that are for nonsymmetric problems) and CGS and BiCG methods that are primarily applicable to symmetric, positive-definite problems [5]. The article also does not describe what kind of direct method is meant by "traditional Gaussian elimination" implemented in the older version of Spice and why it was slower than iterative methods. A better view is given in the article [6] where methods such as GMRES, CGS, and BICGStab are compared to a direct method and tested on several different circuits. Unfortunately, the paper does not present any closer information about tested circuits (only names), and the reader can only guess their sizes and properties. Also, the article does not state whether the algorithms use preconditioning or not. We may note that GMRES works well for nonsymmetric problems, and it is also the most popular method in general. From the performance perspective, with enough memory and good preconditioner it is a very fast and stable method. However, when it comes to real implementation, it shows that the high memory requirements and complexity of computation together with the impossibility to directly access the solution and residual vector at any iterate disadvantages the practical use of this method as a successor to direct sparse solvers. GMRES method accompanied with preconditioner that comes directly from the previously factorized L and U matrices is proposed in the article [7]. There are three methods of computation of preconditioner suggested for GMRES iterative solver. Although it is achieved at the cost of reduced precision (10^{-8}), we must admit that the given algorithm (regardless of used preconditioner) accelerates the computation of transient analysis compared to direct method. As problematic we see that the algorithm can achieve best results only in the case of small changes in variables between transient steps (and low condition number of circuit matrix) ensuring a very low number of iteration cycles of NR algorithm.

II. Scalable Linear Solver

By the term of scalable linear solver is meant an implementation of such computation procedures that optimize analysis performance regardless of problem size. In linear algebra, electronic circuit matrices represent one special group of linear problems. They can be characterized as extremely sparse matrices whose density of nonzero values quadratically decreases with dimension. They are certainly nonsymmetric but permutable to a block of triangular form. In particular

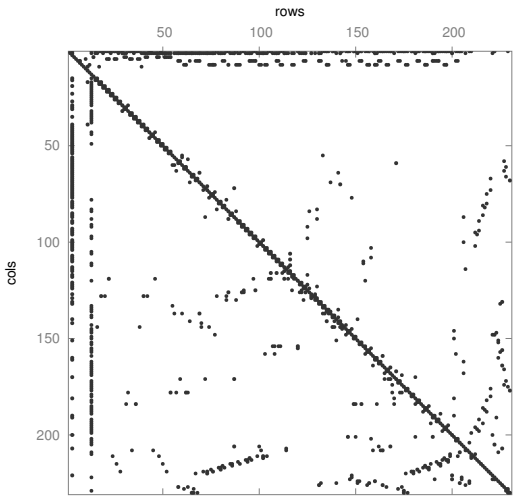


Fig. 1. Visualization of the circuit matrix of the CMOS shifter.

cases, they can be ill-conditioned or temporarily approach singularity. MNF of circuit matrix of CMOS shifter after pivotization produced by program NgSpice is visualized in Fig. 1 to illustrate the situation. The original circuit contains roughly two hundred devices connected by four hundred independent voltage nodes.

Performance of all operation such as factorization, row ordering, and even convergence depends on the matrix properties. Matrix dimension can be considered as the main one. Regardless of used device models, it is dominantly given by the topology of simulated circuit. Based on the dimension of simulation matrix n, we divide circuit simulation into three groups:

$$\begin{cases} \text{Trivial} & n < 50 \\ \text{Standard} & n \in \langle 50, 5000 \rangle \\ \text{Large} & n > 5000 \end{cases}$$

1) In trivial cases, simulation problem is so small that circuit matrix can be considered as almost dense. In that situation, a complicated implementation of the solver of sparse linear systems become meaningless, and can be outperformed by "naive" method. Although relative performance difference may reach 50 percent, in absolute numbers it is clearly below of meaningful resolution.
2) We denoted problems as a standard where the algorithm implemented in Spice for the direct solution of a linear sparse system has no competition. It proves to be a fast solver without extensive routines and able to work with real or complex numbers. It is tough to outperform it even with decent sparse matrix systems such as UMFPACK [8] and SuperLU [9].
3) As large can be considered a linear system compiled from several sub-circuit parts resulting in a matrix of dimension higher than 5000. In the past, this discipline belonged to supercomputers and was taken as a hypothetical possibility. With increasing performance of computers, it is not difficult to perform a simulation of this size in reasonable time. For example, to find a DC solution of the matrix with 41000 non-zero values and dimension 5000 takes less than 0.5s.

The fact that it is tough to outperform the Spice internal direct solver in a region of standard size problems stops to be valid for large scale systems. In that case, iterative numerical methods based on successive approximations gain in importance. In literature GMRES [10] is usually presented as the best choice for an iterative solution of large linear systems. It is applicable to nonsymmetric matrices and leads to the smallest residual with significantly reduced number of iterations. However, we found GMRES problematic from several aspects. The first one, related to nature of the method, is that it is difficult to limit the number of iteration steps in a standard manner. The computation time depends on the limit of available storage and number of restart points. The correct settings of those variables depend on properties of linear system which may vary between Newton-Raphson iteration steps. The fact that the algorithm does not stop iteration at certain point makes this method unsuitable as a competitor to direct solvers. The second aspect relates to relatively high memory requirements of the method. As the most promising proved to be Biconjugate Gradient Stabilized method (BICGStab) [11]. It is often described in literature as a fast and stable method avoiding irregular convergence behaviors [12]. The second promising method turned out to be Quasi-Minimal Residual (QMR) [13] which in some particular cases can be even faster than BICGStab method. As problematic can be taken its slightly worse stability than the one obtained with BICGStab.

TABLE I
Computation Times for Different Parts of NgSpice Direct Solver

Dim./Oper.	802		7138		20098	
	Oper.	Δt (sec)	Oper.	Δt (sec)	Oper.	Δt (sec)
Reorder	4	2.14e-2	4	5.49e-1	4	8.37
LUF	45	1.96e-3	174	0.419	175	2.18
FEBS	45	9.31e-4	174	0.226	175	1.22

A. BiConjugate Gradient Stabilized (BICGStab)

The definition of BiConjugate Gradient Stabilized method (BICGStab) algorithm [11] as it can be found in the literature is usually written in a form of mathematical pseudocode. We find it not very explanatory especially if presented without preceding mathematical derivation. Instead, we refer to simplified implementation in GNU Octave [14] source code. It should be noted that for real implementation that algorithm must be redefined to be able to work with large sparse matrices and also computation with preconditioner needs to be revised. In GNU Octave source code, this is suppressed with a use of backslash operator "\" which performs direct solution of given linear problem. For a more real-world implementation of the algorithm, we recommend looking at the particular method in source codes of Portable, Extensible Toolkit for Scientific computation PETSC [15], [16]. It is evident that every additional iteration requires a certain number of operations that will prolong computation time. It must be reduced so that the method could compete with the direct solver. The optimization can be done with reducing precision or setting maximal iteration number, with more precise initial estimate and with preconditioner.

B. Reduced Precision or Iteration Number

In the article [17] it is proved that for a stable linear solver with a close enough solution, the norm of residual will

be reduced after each step of NR until limiting accuracy is reached. It also states that reduced accuracy of a linear solution may affect the rate of convergence but not limit accuracy. Since the proposed method contains two nested iterative algorithms, it opens a possibility to accelerate computation with a limit of the number of iterations or residual precision of inner method. Mainly for the reason of relatively slow convergence rate of unpreconditioned BICGStab (and all linear solvers in general) in comparison to almost quadratic convergence rate of nonlinear NR, the iteration limit does not provide recognizable performance improvement unless preconditioning is applied. Without it the number of successful iterations of BICGStab method may vary a lot for different linear systems and therefore maximum iteration number must be chosen based on matrix properties. As an indicator may seem to be the matrix condition number efficiently computed by one of the methods published in [18], [19]. They allow estimation of condition number even in the case of highly sparse matrices. Although, it can give us some information about invertibility of a matrix and also absolute achievable precision, it does not prove to be a reliable indicator of iteration numbers for the BICGStab method. A more efficient approach proved to be the use of preconditioner with fixed maximal iteration number.

C. Initial Estimation

Iterative algorithms allow to specify an initial iteration guess. It is a well-known fact that an initial estimate close to the solution rapidly reduces the number of iterations. This works extremely well for NR method and at the first sight it may seem to apply to inner iterative linear solvers. The main issue is given by varying properties of a linear system between iteration steps. Iteration steps computed from Jacobian matrix can transiently become singular or have extremely large condition number. In that situation, we can not expect the solution from the previous iteration to be any close to the subsequent one (it is a big issue for algorithm presented in [7]). Also, linear solvers are more sensitive to wrong initial guess than nonlinear methods when the wrong guess can cause not only increase of required iterations but also a divergence of the method. We illustrate this in Fig. 2, and the example shown here was performed on tridiagonal matrix with dimension 200 defined in (1):

$$a_{i,j} = \begin{cases} 2 & i = j, \\ -1 & i = j - 1, \\ 1 & i = j + 1, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

D. ILU Preconditioner

With growing dimension a linear system gains on the significance of probability of fill-ins. Although reordering and pivotization try to minimize them, it is often not possible to fully avoid them. It implies that to compete with the direct solver we should use a preconditioner that entirely avoids fill-ins but has similar properties. As the best for computation of highly sparse nonsymmetric matrices sorted out Incomplete LU Preconditioner. The primary intention of the ILU is to suppress possible fill-ins maximally. Regardless of modification of the method, it is done simply by dropping

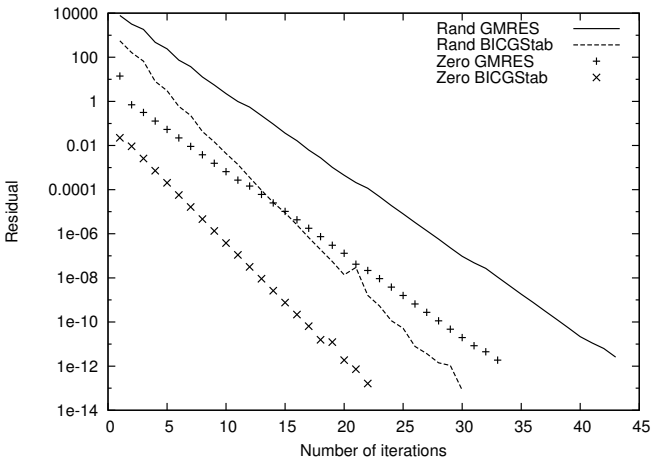


Fig. 2. Effect of starting values on method convergence rate. (Matrix (1).)

them. For subset of nonzero positions in the matrix S, one factorization step can be formally described as

$$A_{i,j} \leftarrow \begin{cases} A_{i,j} - A_{i,k}/A_{k,k}^{-1} a_{k,j} & \text{for } (i,j) \in S, \\ a_{ij} & \text{otherwise,} \end{cases} \tag{2}$$

for each k and for $i, j > k$. It is important to note that the incomplete factorization may fail due to division by zero. It indicates that matrix pivoting can not be avoided even with usage of iterative methods. From an implementation viewpoint, it is not a problem because it is already there and used by the direct method. It should be pointed out that real implementation should be able to work with sparse matrices. Convergence rate of computation of real CMOS Voltage shifter circuit with ILU preconditioned BICGStab method and without it are shown in Fig. 3. The improvement in convergence rate in both methods GMRES and BICGStab is indisputable.

The effect of usage of preconditioner in real simulation performed by NgSpice simulator can be seen in Figs. 4 and 5. It shows average time (on a log scale) needed for the linear solver to compute solution for given iteration step of NR method. The graph was constructed as the average value of hundred repeated analyses of CMOS shifter circuit, but similar results were obtained with other circuits. A very high computation time should be noted in the beginning of simulation with not preconditioned solver as shown in Fig. 4. At this point, OP of the circuit is computed. It is computed with an initial random guess for NR that can even cause singularity of the system. Consequently, it causes longer computation time for the iterative linear solver. This effect is suppressed with a use of preconditioner as shown in Fig. 5.

III. Implementation of Scalable Solver in Spice

The standard simulation procedure needs to be modified to allow scalability of the solver in Spice simulator. All implementations and tests were performed with simulation program NgSpice [20], which is an open source continuation of no longer developed Berkeley's version of Spice. The first attempt to implement iterative linear solvers was made with a source code presented in the book TEMPLATES Iterative Solvers for Linear Systems [12]. Even though the, algorithms presented in the book can be implemented from their source code examples in a very short time, they showed to be not

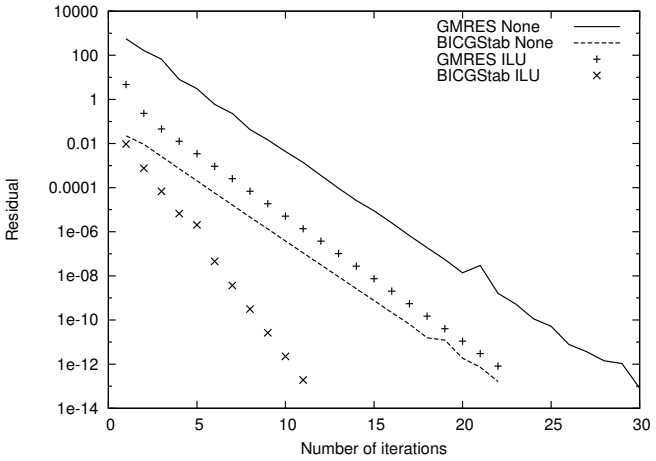


Fig. 3. Effect of used ILU preconditioner on convergence rate. (CMOS voltage shifter.)

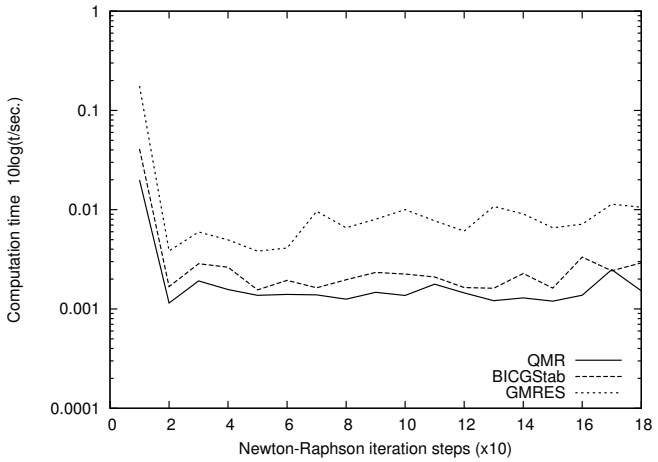


Fig. 4. Required time for each step of transient analysis. (Not preconditioned.)

Algorithm 1 Scalable Linear Solver

```
if Reordering required then
   Pivoting (Markowitz pivot strategy)
end if
if LINSLV=ITER or n > MATSIZE then
   BICGStab Method
   Obtain LINSTOP
end if
if LINSTOP=BREAK or LINSLV=DIRECT then
   Direct method
   Compute residual
end if
if Residual=NaN or Residual=Inf then
   Stop NR
end if
```

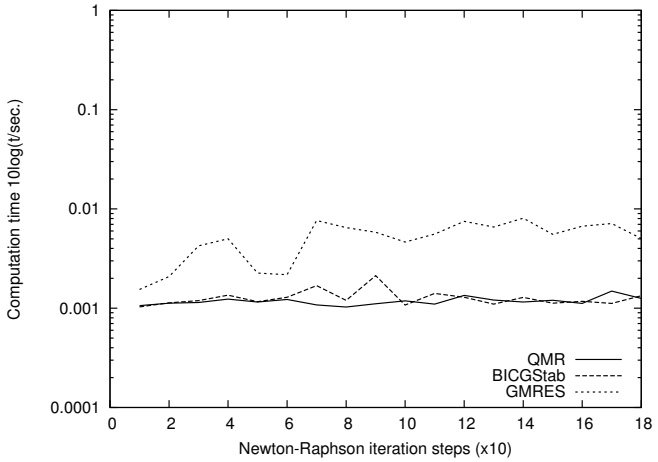


Fig. 5. Required time for each step of transient analysis. (Preconditioned.)

applicable to NgSpice simulation core due to the lack of missing sparse matrix storage system. While they can provide some idea of the difference in performance of algorithms and their convergence properties, it is not possible to compare them with direct methods optimized for sparse matrices implemented in NgSpice. Therefore, we strongly recommend PETSc library [15], [16]. It provides a full variety of iterative linear solvers and even some experimental nonlinear ones. It also provides an interface to external direct solver libraries, such as UMFPACK [8] or SuperLU [9]. We found only one problem with the implementation of incompatible sparse matrix storage between NgSpice and PETSC. Additional translating routines resolved it but for future development, we suggest rewriting of the proprietary NgSpice sparse matrix storage system to one implemented in PETSC. The effect of the modification of original algorithm is shown in Algorithm 1. In the case of small and normal size circuits, it will primarily use the direct linear solver but for large circuits (of a dimension > 5000) it will perform the iterative algorithm. BICGStab method combined with ILU preconditioner proved to be the best iterative method for electronic circuit simulations. Based on matrix size (parameter MATSIZE) it decides whether to use the iterative or direct algorithm. It also allows predefining of the method for given simulation with variable LINSLV. Iteration method by its nature computes with residual. It can be omitted or defined to compute it after a specific number of steps. Our tests proved that avoiding the computation of residual has a very small effect on overall performance when ILU preconditioner is used. On the other hand, the direct method does not compute residual of the solution and therefore it is explicitly specified. Additionally, the algorithm can try to obtain a solution with the direct method based on a current return value LINSTOP of the stopping event. LINSTOP corresponds to PETSC return codes which for positive values indicates convergence and for negative values until -5 method divergency. Lower values than -5 indicates that the process was interrupted by error mostly caused by the occurrence of NaN or INF values in a sparse array. A manual specification of the method should be available to the user through a set of redefinable properties. In a similar manner as used in NgSpice we specify the additional set of simulation parameters:

```
LINSLV # type of linear solver (DIRECT, ITER)
KSPTYPE # linear method BICGStab (Default)
KSPPREC # preconditioner ILU ( Default)
KSPRELTOL # relative stopping tolerance
KSPABSTOL # absolute stopping tolerance
KSPMAXIT # iteration limit
```

Parameter LINSLV predefines linear solver for the analysis and suppresses adaptive selection based on the size of circuit matrix. PETSC library offers a full range of different linear solvers and preconditioners. They are mostly compatible with

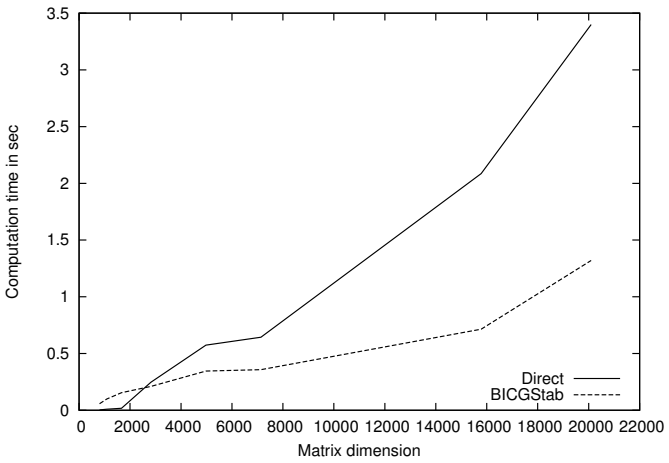


Fig. 6. Performance comparison of iterative BICGStab and direct method implemented in NGspice.

each other. Once the library is integrated to NgSpice, it is possible to use all its methods freely (in our case they can be chosen by definition of parameters KSPTYPE and KSPPREC). We also define parameters restraining maximum number of iterations and the relative and absolute stopping tolerances (KSPRELTOL, KSPABSTOL, KSPMAXIT). In Fig. 6, there is a comparison of our proposed procedure in Algorithm 1 performed by NgSpice with direct method for the solution of linear system. Iterative algorithm BICGStab was implemented with ILU preconditioner. To avoid slow convergence of the method in the case of wrong initial guess of NR method we limited the number of iteration of BICGStab method to 10. The relative precision of the algorithm was set to 10^{-16} and the absolute one to 10^{-50}. Simulation time corresponds to a required time to obtain operation point of given circuit.

IV. Conclusion

It is clear that performance of computer simulation of electrical circuits depends on the implementation of fast and reliable computation procedures. In the article, we proposed adaptive internal solver based on BICGStab for the iterative solution of nonsymmetric linear systems supplemented with incomplete LU factorization in Algorithm 1. Most importantly, the algorithm (based on return values from linear iteration cycle) can automatically switch to direct method or directly reject integration step if linear iteration leads to non-numeral types. It is the extremely efficient operation during simulation of large scale circuit that significantly reduces the number of unnecessary iteration cycles of NR algorithm. The proposed combination of the linear solver and preconditioner greatly improves the convergence rate of the linear method and make it compatible with standard LU solver from circuit matrix dimension up to 5000. For a smaller circuit the iterative solver can obtain a result with improved precision avoiding errors caused by operation with finite floating point precision. In Fig. 6, there is a comparison of time required for computation of OP with the use of direct method and iterative algorithm. In this case, linear solver was manually predefined and therefore no additional "fail-safe" computation was performed. ILU avoids redundant computations and ensures constant the size of the matrix. It allows for the BICGStab iteration, especially during successful iteration cycles of NR algorithm, to obtain solutions much faster than standard direct method. This difference grows with size of the circuit matrix.

Acknowledgments

This paper was supported by the Technology Agency of the Czech Republic under the grant no. TE01020186 and by Czech Technical University grant no. SGS18/079/OHK3/1T/13.

References

[1] T. Nechma and M. Zwolinski, "Parallel sparse matrix solution for circuit simulation on FPGAs," *IEEE Transactions on Computers*, vol. 64, no. 4, pp. 1090–1103, April 2015.

[2] X. Chen, L. Ren, Y. Wang, and H. Yang, "GPU-accelerated sparse LU factorization for circuit simulation with performance modeling," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 26, no. 3, pp. 786–795, March 2015.

[3] N. Kapre and A. DeHon, "SPICE2 : Spatial processors interconnected for concurrent execution for accelerating the SPICE circuit simulator using an FPGA," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 31, no. 1, pp. 9–22, Jan 2012.

[4] M. A. Kleiner and R. N. Afsar, "Steady-state determination for RF circuits using krylov-subspace methods in SPICE," in *Microwave Symposium Digest, 2001 IEEE MTT-S International*, vol. 3, May 2001, pp. 2091–2094 vol.3.

[5] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed. Society for Industrial and Applied Mathematics, 2003. [Online]. Available: http://epubs.siam.org/doi/abs/10.1137/1.9780898718003

[6] M. A. Kleiner and M. N. Afsar, "Determining the steady-state responses in RF circuits using GMRES, CGS and BICGSTAB solution in sSPICE for linux," in *Microwave Symposium Digest. 2000 IEEE MTT-S International*, vol. 1, June 2000, pp. 87–90 vol.1.

[7] Z. Li and C. J. R. Shi, "An efficiently preconditioned GMRES method for fast parasitic-sensitive deep-submicron vlsi circuit simulation," in *Design, Automation and Test in Europe*, March 2005, pp. 752–757 Vol. 2.

[8] T. Davis, *Direct Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, 2006.

[9] X. S. Li, "An overview of SuperLU: Algorithms, implementation, and user interface," vol. 31, no. 3, pp. 302–325, September 2005.

[10] Y. Saad and M. H. Schultz, "GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems," *SIAM Journal on Scientific and Statistical Computing*, vol. 7, no. 3, pp. 856–869, 1986. [Online]. Available: http://dx.doi.org/10.1137/0907058

[11] H. A. van der Vorst, "BI-CGSTAB: A fast and smoothly converging variant of BI-CG for the solution of nonsymmetric linear systems," *SIAM J. Sci. Stat. Comput.*, vol. 13, no. 2, pp. 631–644, Mar. 1992. [Online]. Available: http://dx.doi.org/10.1137/0913035

[12] R. Barrett, M. Berry, T. F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. V. der Vorst, *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods, 2nd Edition*. Philadelphia, PA: SIAM, 1994.

[13] R. W. Freund and N. M. Nachtigal, "QMR: a quasi-minimal residual method for non-hermitian linear systems," *Numerische Mathematik*, vol. 60, no. 1, pp. 315–339, 1991.

[14] J. W. Eaton, D. Bateman, S. Hauberg, , and R. Wehbring, *GNU Octave version 4.0.0 manual: a high-level interactive language for numerical computations*, 2015.

[15] S. Balay, W. D. Gropp, L. C. McInnes, and B. F. Smith, "Efficient management of parallelism in object oriented numerical software libraries," in *Modern Software Tools in Scientific Computing*, E. Arge, A. M. Bruaset, and H. P. Langtangen, Eds. Birkhäuser Press, 1997, pp. 163–202.

[16] S. Balay, S. Abhyankar, M. F. Adams, J. Brown, P. Brune, K. Buschelman, L. Dalcin, V. Eijkhout, W. D. Gropp, D. Kaushik, M. G. Knepley, L. C. McInnes, K. Rupp, B. F. Smith, S. Zampini, H. Zhang, and H. Zhang, "PETSc users manual," Argonne National Laboratory, Tech. Rep. ANL-95/11 - Revision 3.7, 2016. [Online]. Available: http://www.mcs.anl.gov/petsc

[17] F. Tisseur, "Newton's method in floating point arithmetic and iterative refinement of generalized eigenvalue problems," *SIAM J. Matrix Anal. Appl*, vol. 22, pp. 1038–1057, 1999.

[18] E. Deadman, "Estimating the condition number of f(a)b," *Numerical Algorithms*, vol. 70, no. 2, pp. 287–308, 2015. [Online]. Available: http://dx.doi.org/10.1007/s11075-014-9947-4

[19] N. J. Higham and F. Tisseur, "A block algorithm for matrix 1-norm estimation, with an application to 1-norm pseudospectra," *SIAM J. Matrix Anal. Appl*, vol. 21, pp. 1185–1201, 2000.

[20] P. Nenzi and H. Vogt, *Ngspice Users Manual 26*, 2016.

Optic Disc Segmentation Based on Correlation Feature Information

Olawuyi O. Fatoki & Sunday Ojo
Computer Science Dept
Tshwane University of Technology
Pretoria,
South Africa
Email: mayowafatoki@gmail.com; ojoso@tut.ac.za

Pius Adewale Owolawi
Computer Systems Engineering Dept
Tshwane University of Technology
Pretoria,
South Africa
Email: owolawipa@tut.ac.za

Temitope Mapayi
Electrical Engineering Dept
Mangosuthu University of Technology
Umlazi, Durban,
South Africa.
Email: mapayi.temitope@mut.ac.za

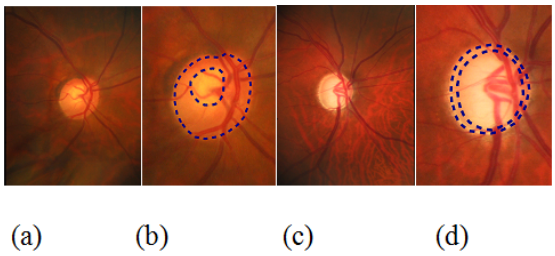
***Abstract*—Glaucoma, one of the major causes of blindness, has been identified as a disease that causes the degeneration of the optic disc. An highly accurate automated detection of the optic disc (OD) has however been noted to be of great importance for the efficient diagnosis of the disease. This paper presents a study on an unsupervised approach usage of Haralick Correlation texture feature for the segmentation of optic disc in colored fundus retinal images. The grayscale and green channel of the colored fundus image are investigated. When compared with some methods in the literature, the experimental study of the proposed method achieved very high average accuracy rates of 98.59% and 98.36% using grayscale and green channel of the colored fundus image respectively on DRIVE database.**

Keywords*—Correlation, Haralick Feature, Glaucoma, Optic disc, Segmentation, image processing.

I. Introduction

Glaucoma has been identified as the second leading cause of blindness with approximately 60 million cases reported globally in 2010 [1]. Report has also indicated that an estimation of about 80 million people are likely to suffer from glaucoma by 2020 [1]. It is an ailment that causes impairment to the optic disc and the nerves that convey information from the eye to the brain. It has been related with growth in the IOP of the eye. Glaucoma largely causes progressive damage to outer vision due to irreparable damage in the optic nerve and optics nerve head (ONH). This ailment can lead to permanent blindness if not treated early. Hence, the early diagnosis of glaucoma is very important for a proper management of the disease [2].

There are basically two different categories of glaucoma and they are known as open-angle glaucoma and angle-closure glaucoma. Open-angle glaucoma, which is also known as wide-angle or chronic simple glaucoma, is the most common type of glaucoma. The structures of the eye seem normal, but the pressure within the eye rises gradually, and vision loss progresses over a period of time. Angle-closure glaucoma, on the other hand, is the chronic angle-closure or narrow-angle or acute glaucoma, the pressure inside the eye rises suddenly with this type of glaucoma and immediate medical intervention is very paramount to preserve vision. Fig.1 shows an example of healthy and glaucoma-suspicious eye. Fig. 1(a) and 1(b) show the optic disc images of healthy eye while Fig. 1(c) and 1(d) show the optic disc images of Glaucoma-suspicious eye from RIM-ONE v.3 [3] database. Fig. 1(b) and Fig 1(d) contain enlarged areas of optic discs and the borders of optic discs are indicating by outer dashed line and optics cup borders by inner dashed line.



(a) (b) (c) (d)

Fig. 1: (a) and (b) show the optic disc images of healthy eye (c) and (d) show the optic disc images of Glaucoma-suspicious eye [3] database.

The optic disc has been acknowledged to be of good help in shedding light on the relationship between the optic nerve cupping and loss of visual field in glaucoma [4]. The automated detection of the optic disc has also been identified as a tool that will be of great usefulness for the diagnosis of Glaucoma in countries that significantly lack qualified physicians [5], [6].

Several methods have been proposed in the literature for the automated segmentation of optic disc and significant progress has been achieved. An optic disc and optic cup segmentation method was proposed in [7]. Park et al.,[8] implemented an optic disc segmentation in retinal images using thresholding and Hough transform. Tjandrasa et al. [9] combined Hough transform and active contours for the segmentation of the optic disc. An optic disc segmentation technique based on color difference and vessel bends information was implemented in [10]. An adaptive thresholding techniques for the segmentation of optic disc and optic cup was implemented in [11].

A supervised learning approach based on k-nearest neighbor was implemented by Septiarini et al., [12]. Fu et al., [13] implemented one-stage multilabel using deep learning based architecture for the segmentation of optic disc and optic cup.

Although some of these previously proposed methods have achieved progress using different image processing techniques for the segmentation of optic disc, some problems remain unresolved. The extraction of relevant image features is very important in the process of OD segmentation and in most

cases, disappearing of the blood vessel in the OD leads to poor OD segmentation [7]. Another reason for poor OD segmentation is the misclassification of OD pixels due to some problems at the preprocessing phase during the removal of noise from the fundus image [11], [12]. Since the efficient segmentation of the optic disc is one of the most critical task in automatic diagnosis of patients for glaucoma disease, good segmentation techniques are highly needed to produce good (OD) segmentation result. One of the important cues applied for object segmentation in image analysis is texture. This research presents a study on an unsupervised approach usage of correlation Haralick feature [14] for the segmentation of optic disc in retinal fundus images.

The remaining part of this paper is presented as follow: The methods and techniques used in this study are presented in Section II, experimental set-up, results and discussion are presented in Section III and the conclusion is drawn in Section IV.

II. Methods & Techniques

The automatic segmentation of optic disc in retinal fundus images is presented in this section. The fundus images first undergo some preprocessing techniques for some image enhancement and noise removal. This is followed by the segmentation of the optic disc based on Haralick texture correlation information using an unsupervised approach where the texture information is used to compute a local adaptive threshold. Lastly, the remaining unwanted noise after segmentation are removed in the postprocessing phase.

A. Fundus Image Preprocessing

This study investigates the use of grayscale image and the green channel of the RGB colored fundus image. At the preprocessing phase, the extraction of the green channel of the colored fundus image is performed. The grayscale of the colored fundus image is also computed. A median filter of a window size *n*n* is then applied for the enhancement of the image as:

$$Im^{convolved}(i,j) = Med(x,y) * Im^{gr}_{n*n}(x,y) \tag{1}$$

where n=6, $Im^{gr}(x,y)$ is the extracted green channel or the grayscale of the fundus image, $Im^{convolved}(i,j)$ is the convolved image and the mask *Med(x,y)* used for the convolution is the median filter.

This is followed by a top-hat transform based on a disk-shaped structure element that is used for the contrast enhancement of the fundus images. The top-hat transform is computed by subtracting the opening of the result obtained in equation 1 from $Im^{convolved}(i,j)$ (i.e. opening of $Im^{convolved}(i,j)$).

$$Im^{tophat}(i,j) = Im^{convolved}(i,j) - (Im^{convolved}(i,j) \circ B(u,v)) \tag{2}$$

where B(u, v) is the disk-shaped structure element.

B. Optic Disc Segmentation

Texture is an important characteristic used in identifying regions of interest in an image and has also been identified as a key component of human visual perception. Unlike colour, texture occurs over a region rather than at a point. Texture has qualities such as periodicity and scale; and it can be described in terms of direction, coarseness, contrast [15].

Texture analysis has played an important role in several medical applications. In general, the applications involve the automatic extraction of features from the image, which is then used for a variety of classification tasks. Applications of texture analysis may include segmentation of different anatomical structures in the human body. Gray Level Co-occurrence Matrices (GLCM), introduced by Haralick [14], is a powerful technique for measuring texture features. Different Haralick features [14] have been implemented for supervised and unsupervised segmentation methods in the literature.

The implementation of Haralick texture-based methods using unsupervised approach where adaptive local thresholds are used for image segmentation in the literature [16], [17], [18] have achieved promising results. Six of the fourteen Haralick features, namely entropy, energy, variance, contrast, local homogeneity and correlation, have been considered to be very relevant in [19], [20]. This makes correlation, one of the six very relevant features, an essential feature to investigate in this study for the segmentation optic disc in retinal images.

Given that i is the gray-level intensity of pixel (x,y) in the fundus image, the GLCM of the fundus image is computed as:

$$C_{i,j} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} (P\{V(x,y) = i \,\&\, V(x \pm d\Phi_1,\ y \pm d\Phi_2) = j\}) \tag{3}$$

where P is defined as

$$P(x) \quad = \quad \begin{cases} 1 & \text{if x is true} \\ 0 & \text{Otherwise} \end{cases} \tag{4}$$

The correlation feature is computed as:

$$Correl = \sum_i \sum_j Pd(i,j) \frac{(i-\mu_x)(j-\mu_y)}{\sigma_x \sigma_y} \tag{5}$$

where $\mu_x \mu_y$ and $\sigma_x \sigma_y$ are the mean and standard deviation respectively and are computed in equations 6 to 9 below as

$$\mu_x = \sum_i \sum_j iPd(i,j) \tag{6}$$

$$\mu_y = \sum_i \sum_j jPd(i,j) \tag{7}$$

$$\sigma_x = \sqrt{\sum_i \sum_j (i-\mu_x)^2 Pd(i,j)} \tag{8}$$

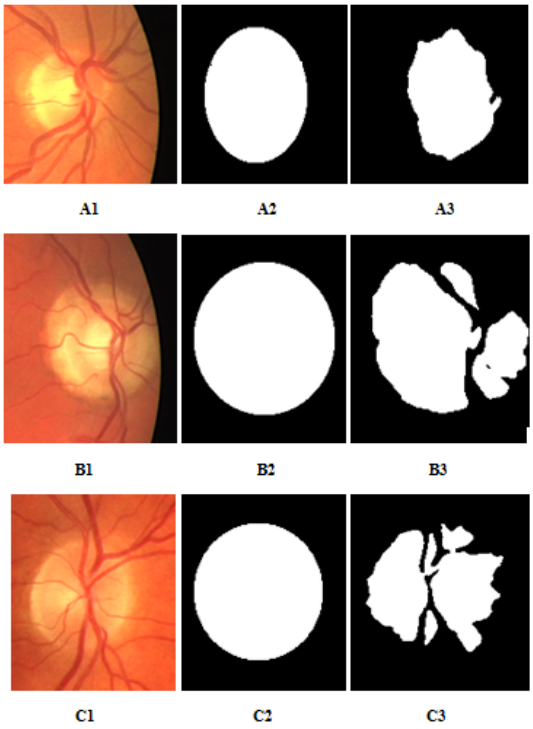


Fig. 2: A1, B1 & C1 shows optic discs in colored fundus images, A2, B2 & C2 show the ground truth and A3, B3 & C3 show the result of the proposed segmentation Method

$$\sigma_y = \sqrt{\sum_i \sum_j (i - \mu_y)^2 Pd(i,j)} \tag{9}$$

A multi-scale feature metric of the correlation feature across the distance *'d'* and relative orientation *'Φ'* is computed as:

$$Cor^1 = (Cor_{ij}), 1 \leq i,j \leq 4 \tag{10}$$

where

$$Cor_{ij} = Cor_{(d_i,\Phi_j)}, 1 \leq i,j \leq 4 \tag{11}$$

such that Φ_1= 0^o, Φ_2= 45^o, Φ_3= 90^o and Φ_4 = 135^o, with distances $(d_i)_{i=1,\ldots,4}$. The correlation texture feature is investigated in this study across the distances $(d_i)_{i=1,\ldots,4}$ over the four orientations (i.e. 0^o , 45^o, 90^o and 135^o) to compute the local adaptive threshold used for the optic disc segmentation implemented in this study.

The local adaptive threshold based on correlation information used for the optic disc segmentation is then computed as:

$$OD_{Th} = \frac{\max_{1\leq i\leq 4} \parallel \max_{1\leq j\leq 4}(Cor_{ij}) - \min_{1\leq j\leq 4}(Cor_{ij}) \parallel}{(4*4)} \tag{12}$$

Given that

$$F(x,y) = Im^{tophat}(i,j) \tag{13}$$

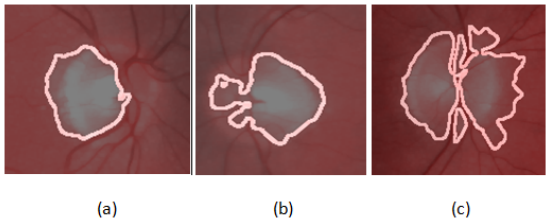


Fig. 3: Optic disc Segmentation in Colored Fundus Images

TABLE I: Comparison of the Performance of the Proposed Method with Other Segmentation Methods on DRIVE Database

Method	Avg. Accuracy
Abdullah et al. [22]	96.72%
Abdali-Mohammadi & Poorshamam [23]	97.5%
Lupascu et al., [24]	95%
Morales et al. [25] (PCA)	99.03%
Morales et al., [25] (Red Channel)	89.79%
Proposed Method (Gray Intensity)	**98.59%**
Proposed Method (Green Channel)	**98.36%**

The segmented optic disc is given below as:

$$OD_{seg} = \begin{cases} 0, & \text{if } F(x,y) \leq OD_{Th} \\ 1, & \text{otherwise} \end{cases} \tag{14}$$

C. Post-Processing

The postprocessing phase involves the removal of the remaining noise using morphological opening. A morphological opening operation utilize an erosion operation which is followed by a dilation operation. The retinal fundus image mask is then subtracted from the image obtained after applying morphological opening to get the final result of the optic disc segmentation as shown in Figs. 2 & 3.

III. Experimental Results and Discussion

The experiments conducted in this study were performed using MATLAB 2014a. The method proposed in this paper was evaluated using the publicly available DRIVE [21] database that contain retinal fundus images. The database contains 40 fundus images which were captured using Canon CR5 camera of a 24-bit gray-scale resolution, having the spatial resolution of 565 x 584 pixels. The 40 images are made up of training set containing twenty images and testing set containing twenty images. The provided ground truth of the optic discs on the DRIVE database are used for the performance evaluation of the method proposed in this study. Accuracy metric is the performance measure used in this paper and it is computed as:

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{15}$$

where TN is true negative, TP is true positive, FN is false negative and FP is false positive.

Table 1 presents the performance comparison of the proposed method using both grayscale and green channel of the RGB. The average accuracy rate measure presented in Table 1 is computed by dividing the total accuracy rate by

the total number of the segmented optic disc images. It is also shown in Table 1 that the grayscale achieved a slightly higher accuracy rate of 98.59% when compared with the green channel that achieved an accuracy rate of 98.36%. The proposed method using both grayscale and green channel achieved higher accuracy rates of 98.59% and 98.36% when compared with the methods proposed by Abdali-Mohammadi & Poorshamam [23], Lupascu et al.,[24] and Morales et al., [25] (Red Channel). The method proposed by Morales et al. [25] based on principal component analysis (PCA) have a higher accuracy rate of 99.03% when compared to the proposed method using both grayscale and green channel that achieved the accuracy rates of 98.59% and 98.36% respectively on DRIVE.

IV. Conclusion

Glaucoma has been identified as a disease that degenerates the optic disc, and if left untreated will eventually lead to blindness. It has also been noted that an highly accurate automated detection of the optic disc in fundus images is required for the efficient diagnosis of the disease. This paper has presented a study on an unsupervised approach usage of correlation Haralick feature for the segmentation of optic disc. The proposed method investigated the use of grayscale and green channel the colored fundus image and achieved very promising results with accuracy rates of 98.59% and 98.36% respectively. Our future work shall extend the evaluation of the proposed method on other publicly available dataset and improve on the method.

References

[1] H. A. Quigley and A. T. Broman, "The number of people with glaucoma worldwide in 2010 and 2020," *British journal of ophthalmology*, vol. 90, no. 3, pp. 262–267, 2006.

[2] C. Costagliola, R. Dell'Omo, M. R. Romano, M. Rinaldi, L. Zeppa, and F. Parmeggiani, "Pharmacotherapy of intraocular pressure: part i. parasympathomimetic, sympathomimetic and sympatholytics," *Expert opinion on Pharmacotherapy*, vol. 10, no. 16, pp. 2663–2677, 2009.

[3] F. Fumero, S. Alayón, J. L. Sanchez, J. Sigut, and M. Gonzalez-Hernandez, "Rim-one: An open retinal image database for optic nerve evaluation," in *2011 24th international symposium on computer-based medical systems (CBMS)*. IEEE, 2011, pp. 1–6.

[4] C. Duanggate, B. Uyyanonvara, S. S. Makhanov, S. Barman, and T. Williamson, "Parameter-free optic disc detection," *Computerized Medical Imaging and Graphics*, vol. 35, no. 1, pp. 51–63, 2011.

[5] A. Bastawrous, H. K. Rono, I. A. Livingstone, H. A. Weiss, S. Jordan, H. Kuper, and M. J. Burton, "Development and validation of a smartphone-based visual acuity test (peek acuity) for clinical practice and community-based fieldwork," *JAMA ophthalmology*, vol. 133, no. 8, pp. 930–937, 2015.

[6] V. Lodhia, S. Karanja, S. Lees, and A. Bastawrous, "Acceptability, usability, and views on deployment of peek, a mobile phone mhealth intervention for eye care in kenya: qualitative study," *JMIR mHealth and uHealth*, vol. 4, no. 2, 2016.

[7] N. Inoue, K. Yanashima, K. Magatani, and T. Kurihara, "Development of a simple diagnostic method for the glaucoma using ocular fundus pictures." in *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the*. IEEE, 2006, pp. 3355–3358.

[8] M. Park, J. S. Jin, and S. Luo, "Locating the optic disc in retinal images," in *Computer Graphics, Imaging and Visualisation, 2006 International Conference on*. IEEE, 2006, pp. 141–145.

[9] H. Tjandrasa, A. Wijayanti, and N. Suciati, "Optic nerve head segmentation using hough transform and active contours," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 10, no. 3, pp. 531–536, 2012.

[10] M. Hu, C. Zhu, X. Li, and Y. Xu, "Optic cup segmentation from fundus images for glaucoma diagnosis," *Bioengineered*, vol. 8, no. 1, pp. 21–28, 2017.

[11] A. Agarwal, S. Gulia, S. Chaudhary, M. K. Dutta, R. Burget, and K. Riha, "Automatic glaucoma detection using adaptive threshold based technique in fundus image," in *Telecommunications and Signal Processing (TSP), 2015 38th International Conference on*. IEEE, 2015, pp. 416–420.

[12] A. Septiarini, D. M. Khairina, A. H. Kridalaksana, and H. Hamdani, "Automatic glaucoma detection method applying a statistical approach to fundus images," *Healthcare informatics research*, vol. 24, no. 1, pp. 53–60, 2018.

[13] H. Fu, J. Cheng, Y. Xu, D. W. K. Wong, J. Liu, and X. Cao, "Joint optic disc and cup segmentation based on multi-label deep network and polar transformation," *arXiv preprint arXiv:1801.00926*, 2018.

[14] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *Systems, Man and Cybernetics, IEEE Transactions on*, no. 6, pp. 610–621, 1973.

[15] H. Tamura, S. Mori, and T. Yamawaki, "Textural features corresponding to visual perception," *IEEE Transactions on Systems, man, and cybernetics*, vol. 8, no. 6, pp. 460–473, 1978.

[16] T. Mapayi, S. Viriri, and J.-R. Tapamo, "A new adaptive thresholding technique for retinal vessel segmentation based on local homogeneity information," in *Image and Signal Processing*. Springer, 2014, pp. 558–567.

[17] T. Mapayi, J.-R. Tapamo, and S. Viriri, "Retinal vessel segmentation: A comparative study of fuzzy c-means and sum-entropy information on phase congruency," *International Journal of Advanced Robotic Systems*, vol. 12, no. 1, p. 11, 2015.

[18] T. Mapayi and J.-R. Tapamo, "Sahf: Unsupervised texture-based multiscale with multicolor method for retinal vessel delineation," in *International Symposium on Visual Computing*. Springer, 2016, pp. 639–648.

[19] T. Mapayi, S. Viriri, and J.-R. Tapamo, "Adaptive thresholding technique for retinal vessel segmentation based on glcm-energy information," *Computational and mathematical methods in medicine*, vol. 2015, 2015.

[20] S. Di Cataldo, E. Ficarra, A. Acquaviva, and E. Macii, "Segmentation of nuclei in cancer tissue images: Contrasting active contours with morphology-based approach," in *BioInformatics and BioEngineering, 2008. BIBE 2008. 8th IEEE International Conference on*. IEEE, 2008, pp. 1–6.

[21] "Research section, digital retinal image for vessel extraction (drive) database. utrecht, the netherlands, univ. med. center utrecht, image sci. inst. [online]. available: http://www.isi.uu.nl/Re-search/Databases/ DRIVE."

[22] M. Abdullah, M. M. Fraz, and S. A. Barman, "Localization and segmentation of optic disc in retinal images using circular hough transform and grow-cut algorithm," *PeerJ*, vol. 4, p. e2003, 2016.

[23] F. Abdali-Mohammadi and A. Poorshamam, "Automatic optic disc center and boundary detection in color fundus images," *Journal of AI and Data Mining*, vol. 6, no. 1, pp. 35–46, 2018.

[24] C. A. Lupascu, D. Tegolo, and L. Di Rosa, "Automated detection of optic disc location in retinal images," in *Computer-Based Medical Systems, 2008. CBMS'08. 21st IEEE International Symposium on*. IEEE, 2008, pp. 17–22.

[25] S. Morales, V. Naranjo, J. Angulo, and M. Alcañiz, "Automatic detection of optic disc based on pca and mathematical morphology," *IEEE transactions on medical imaging*, vol. 32, no. 4, pp. 786–796, 2013.

Improving Higher Education Institutions' Innovative Teaching and Learning in Tanzania Through Information Communications Technology Tools

Godfrey Isaac Mwandosya
ICT and Mathematics Department
College of Business Education (CBE)
Dar es Salaam, Tanzania
g.mwandosya@cbe.ac.tz

Calkin Suero Montero
School of Computing
University of Eastern Finland
Joensuu, Finland
calkin.montero@uef.fi

Esther Rosinner Mbise
Marketing Department
College of Business Education
Dar es Salaam, Tanzania
ermbise@gmail.com

Edda Tandi Lwoga
Academic Research and Consultancy
College of Business Education
Dar es Salaam, Tanzania
t.lwoga@cbe.ac.tz

Respickius Casmir
Undergraduate Studies
College of Business Education
Dar es Salaam, Tanzania
r.casmir@cbe.ac.tz

Moses Nkundwe Mwasaga
Computer Studies Department
Dar es Salaam Institute of Technology
Dar es Salaam, Tanzania
moses.nkundwe@gmail.com

***Abstract*—The application of information communications technologies (ICTs) in higher education institutions has created opportunities for improving teaching and learning. One such avenue is innovation. However, in the Tanzanian HEIs environment little is known on the extent ICTs tools are utilized for innovative teaching and learning. The aim of this study was to find the extent ICT infrastructure, teachers' skills, and readiness influence innovative teaching and learning. Guided by the Activity Theory, the sample for the study was 80 teachers and Information Technology (IT) staff from three HEIs. Data was collected through an administered survey questionnaire and interviews. The results indicate that the infrastructure is adequate and teachers have skills and are ready to use ICT tools in classrooms. However, wireless technology is inadequate in all the three HEIs. Furthermore, the results show that innovative curriculums and collaborative innovation projects among the different HEIs are factors for improving innovative teaching and learning.**

Keywords—innovative teaching and learning, innovation, ICT tools, HEIs, Tanzania

I. INTRODUCTION

The use of information communication technology (ICT) in higher education institutions (HEIs) in Tanzania as a pedagogical tool is gaining momentum [1]. Different studies in Tanzania and other regions in Africa have reported on the integration of ICT into the teaching and learning process in classrooms [2, 1, 3, 4, 5]. The government of Tanzania has put in place an ICT policy to guide different sectors on how best ICT should be used in enhancing the sector's functionalities [6]. Similarly, the education training policy has been formulated [7]. The emphasis elaborated in the policies is on using technologies to improve a sense of innovations in different sectors in line with the Tanzanian 2025 development vision [8]. One of the sectors in question is the education sector, and specifically, higher education institutions. In enhancing the education sector, teaching and learning need to be improved [9] and that ICT tools are one of the important ingredients [10]. This leads to innovation [11] in the sphere of education whereby through ICT tools the traditional teacher-centered way of delivering education is replaced by student-centered collaborative learning enabling flexible access of learning and communication [12]. Also, teaching innovation in HEIs has a positive effect on the understanding of the knowledge and skills for the students [13]. Furthermore, with the ICT infrastructure being available, and skills in using the ICT tools for teaching and learning process in classrooms, also the teachers need to be willing and ready to change to innovative teaching through technology [14]. According to [2], innovations in the sphere of teaching and learning need to take cognizance of and address factors in the broader national and institutional contexts, as well as those relating to the readiness of individual staff to accept change. The successful implementation and usage of ICT tools in HEIs need the infrastructure to be there and also teachers to have skills and capacity to enable them to integrate ICT in classrooms for teaching and learning [15]. The use of ICT tools by teachers in classrooms will enable students to download, organize, access, use, and share educational materials through technologies [5]. In this case, the teachers need to have the required skills in ICT. UNESCO ICT competency framework for teachers [16] reported basic characteristics of professional teachers as (1) ability to learn using ICT, (2) ability to solve complex real-world problems using ICTs and (3) ability to create new knowledge using ICTs. However, in Tanzanian HEIs little is known on the extent of teachers' ICT skills, the adequacy of ICT infrastructure, and the willingness of the teachers to integrate ICT tools to improve innovative teaching and learning.

This study is guided by the activity theory, a philosophical and cross-disciplinary framework for studying different forms of human practices as development process, both individual and social levels interlinked at the same time [17]. The objectives of this study are, therefore (a) to explore the characteristics and adequacy of the current ICT infrastructure and resources to enable ICT tools integration in the teaching and learning; (b) to investigate the skills and readiness of teachers of HEIs in Tanzania to integrate ICT tools in order to improve innovative teaching and learning; and (c) to explore from the perspectives of teachers and IT staff how best to improve innovative teaching and learning in HEIs in Tanzania. In order to address the objectives, this study sought to answer the following research questions:

1. To what extent is the current ICT infrastructure and resources adequate to enable ICT tools integration in the

teaching and learning in higher education institutions in Tanzania?

2. What are the ICT skills and readiness of teachers of higher education institutions in Tanzania in integrating ICT tools for improving innovative teaching and learning?
3. What are the teachers' views on the different ways for the improvement of innovative teaching and learning in higher learning institutions in Tanzania?

This study is geared towards providing insights for teachers in HEIs in Tanzania in embracing information and communication technology use, and at the same time acts as one of the planning tools for innovative teaching in other higher education institutions emerging economies like Tanzania.

II. BACKGROUND OF THE STUDY

A. The Activity Theory

To address the three research objectives and research questions, our study was guided by the Activity Theory. According to Hashim and Jones [18], activity theory sees the integration of technology as tools which mediate social action. These tools, or artifacts, include instruments, signs, language, machines, and computers. The activity theory was first born within the Soviet psychology, but there is an emerging multidisciplinary and international community of scientific thought united by the central category of activity – a community reaching far beyond the original background. According to Kuutii [17], the structure of an activity comprises four elements; *tool, subject, object,* and the *outcome* realized after the transformation process (see figure 1 below).

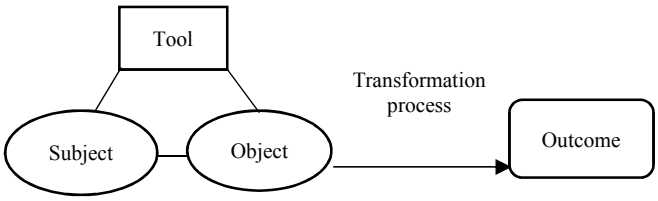


Fig. 1. The structure of an activity adapted from Kuutti [17].

B. ICT tools Integration in Education

Researches point out a number of studies related to the applications, planning, and use of ICT tools in the education sector in higher education in Tanzania. A study by Kihoza et al., [1], assessed the classroom ICTs integration and opportunities in three HEIs in Tanzania using the technological pedagogical and content knowledge (TPACK) and substitute augmentation modification and redefinition (SAMR) models. Mtebe et al., [2] report on the successful implementation of e-learning at the University of Dar es Salaam (UDSM) whereby through the Centre for visual learning, UDSM was able to support conventional face-to-face programmes and ICT-mediated distance programmes. Furthermore, UDSM was able to shift the LMS from Blackboard to Moodle. A study by Mtega et al., [19] narrated a case study at the Sokoine University of Agriculture showing a successful integration of technology in the teaching and learning environment. Kafyulilo et al. [20] in their study, they looked into equipping pre-service teachers with appropriate skills for integrating technology to teach science and mathematics in Tanzania. The study applied the TPACK model whereby the results showed that perceived knowledge and skills of integrating technology in teaching, showed significant changes in technology-related components of TPACK. The integration and use of ICT tools in higher education lead to improvement of innovative teaching and learning [21].

C. Innovative Teaching and Learning

Advancement of technology changes the way people work. For example, the advent of the Web as a social technology has created opportunities for the creation of informal learning environments, which have the potential for innovative methodologies in learning, teaching, and assessment Kivunja [21]. The web 2.0 tools are found to have an impact in the education sector whereby these tools are changing practices, practices that involve shifting from the web as a content repository and information retrieval mechanism to a web that enables more educational and social mediation and user generation of content [22].

D. Improving Teaching and Learning

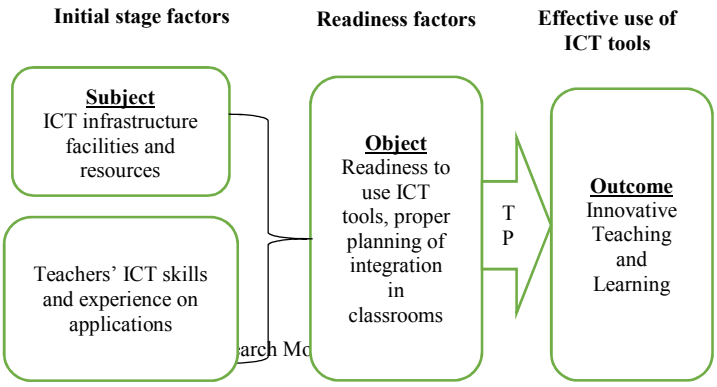
The widespread availability of ICTs and the related services have made it possible for the improvement of teaching and learning in HEIs. The design and development of educational tools to aid teachers and students in the delivery and learning respectively are emphasized in HEIs for example, [23]. In South Africa, a mobile tools and service platform (MOBILED) intended for enhancing formal and informal learning and empowering students within and outside the school premises was established and proved to be of help in enhancing innovative teaching and learning [24]. A study by [25] had used the ICT and digital tools to enhance students learning environmental education.

The importance of Web 2.0 tools in supporting teaching and learning in higher education context in Africa have extensively been described also [26]. Furthermore, based on web 2.0 tools, Conole and Alevizou [22], asserts that the technological environment within which modern education is offered is becoming increasingly complex; it offers new possibilities but also gives rise to challenges. Some of the reported challenges were the low awareness on e-learning and Web 2.0 issues and change of mindset to e-learning activities in Tanzania [26]. Furthermore, most of the faculty members were found to be reluctant to use ICT for teaching purpose [26]. The importance of ICT infrastructure in HEIs, ICT tools, teachers' ICT skills and readiness to use ICT tools in teaching have prompted this study to design a research model described below

E. The Study Research Model

In our study, we investigate the contribution of ICT tools in enhancing innovative teaching and learning through ICT infrastructure, teachers' skills and readiness to use ICT tools. We examined the ICT infrastructure facilities and resources, teachers' skills and experience on applications in place of the subject and are considered as initial factors variable; the ICT tools in place of tools, teachers' readiness, the transformation process (TP) involve effective use of ICT tools and proper

planning of integration in classrooms in place of the object considered as readiness factors variable. Finally, the outcome is the innovative teaching and learning. From this background, it is theorized through our research model (figure 2) that the availability of ICT infrastructure, teachers' ICT skills, and experience will lead to innovative teaching and learning.; *initial stage factors*, readiness *factors*, and *effective use of ICT tools* in improving the innovative teaching and learning in higher education institutions in Tanzania. For a successful implementation and use of ICT tools, the ICT infrastructure needs to be in place. Tedre et. al., [15] indicated that the infrastructure and human capacity are all essential in the implementation of electronic learning (e-learning) at the Tumaini University Iringa (now University of Iringa), Tanzania.



F. *Initial stage factors*

The initial stage factors in our conceptual framework have two elements; the ICT infrastructure facilities and resources, and the teachers' ICT skills and experience with ICT tools' applications. The skills of teachers in handling, using, and experience in using ICT tools is of importance and as Ertmer and Ottenbreit-Leftwich [27] reminds, that teachers should be knowledgeable in science, technology, and pedagogical aspects.

G. *Readiness factors*

Teachers' willingness to use ICT tools is an important factor to consider. In our study we emphasize that motivating teachers to use ICT tools has an impact, as also put forward by Makki and Makki [28], that since the ultimate goal of educational technology instruction is to influence pre-service teachers' ability and willingness to use technology effectively in their teaching careers, it is worthwhile to first consider factors associated with teachers' instructional use of computers. A study by Mwalongo [29] found that teachers' training in ICT is vital for the readiness, future conception and uses of computers for the teaching-learning process. Tedre et al., [15] narrate on planning, preparation in terms of human capacity and infrastructure for the e-learning to effectively take place at the Tumaini University, Iringa University College.

H. *Effective use of ICT tools*

The good planning of integrating ICT tools in classrooms goes hand-in-hand with the proper use of the tools. It must go with consideration to the kind of tools to be applied at the right time, at the right place and at the right environment to cope with the pedagogical content and the level of students.

III. METHODS

The study adopted a mixed approach research design. Three higher education institutions in Tanzania was the population of study. These were: Tumaini University Dar es Salaam College (TUDARco), the College of Business Education (CBE), and the Dar es Salaam Institute of Technology (DIT). All are located in Dar es Salaam. All of the three higher education institutions have campuses in other regions in Tanzania and are known for their academic excellence (DIT for technology-related education[1]), (CBE for business-related education [2]), and (TUDARco for library information services educatio[3]). All of these higher education institutions own basic ICT infrastructures such as a local area network (LAN), Internet, computers, and mobile technology. These form the basis for ICT tools application and establishment of e-learning. Therefore, the selected HEIs provided a good setting for the study.

A. *Study Design and Sampling*

Survey strategy was used to identify and assess the initial stage factors, readiness factors leading to the effective use of the ICT tools for improving the innovative teaching and learning in HEIs in Tanzania. Interviews were carried out to 9 teachers (3 from each HEI) purposely selected out of 18 who use ICT tools in teaching. A survey questionnaire was distributed to 80 teachers and information technology (IT) staff in the three HEIs under study.

B. *Data Collection*

The study was conducted during the last semester of the Academic year 2017/2018 from June to August 2018. Data was collected using a questionnaire divided into four sections seeking: staff background information, skills of teachers to integrate and use ICT tools, and the readiness of teachers to use ICT tools for innovative teaching in classrooms. The interviews involving three teachers from each college/institute sought how and in which ways the innovative teaching and learning in HEIs in Tanzania using the ICT tools could be improved.

C. *Data Analysis*

Collected data was first categorized into different themes related to variables in the conceptual framework. The initial stage factors and readiness factors were analyzed using the statistical package for social sciences (SPSS). In addition, descriptive statistics involved frequencies and percentages. The qualitative data were analyzed through the content analysis to gain in-depth information and views on how best to improve innovative teaching and learning in HEIs in Tanzania.

[1] www.dit.ac.tz
[2] www.cbe.ac.tz
[3] www.tudarco.ac.tz

IV. RESULTS

A. Demographic Characteristic of Respondents

Teachers who participated were 80 in total, 40 (50%) from the College of Business Education, 20 (25%) from Tumaini Dar es Salaam College, 20 (25%) from Dar es Salaam Institute of Technology.

Out 80 teachers 59 (74%) were males and 21 (26%) females. In total, the teaching staffs were 61(76%) and 19 (24%) IT staff.

B. Interview Results - Outcome

The interviews were done with three teachers from each of the colleges. The interviews were in response to a question that wanted to obtain teachers views on how best to improve innovative teaching and learning through ICT tools? The interviews produced three themes; ICT infrastructure improvement, Curriculum change to suit innovation, and teachers' readiness as crucial for innovative teaching (see table 1).

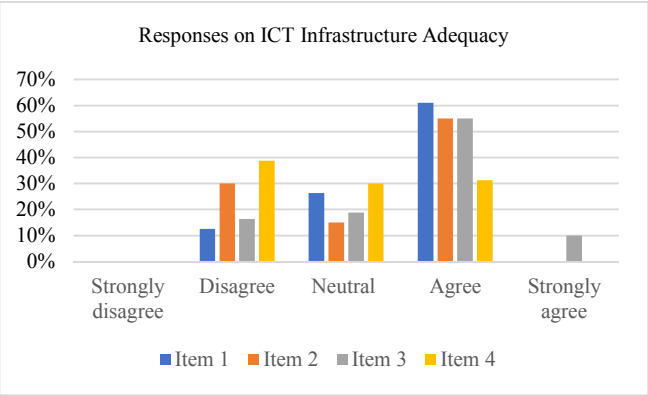
TABLE 1. RESPONSES TO IMPROVING INNOVATIVE TEACHING AND LEARNING THROUGH ICT TOOLS

Question	CBE response	TUDARco response	DIT response
How best to improve innovative teaching and learning through ICT tools?	1. Infrastructure improvement 2. Special classrooms for IT training 3. Training on ICT tools' usage 4. Collaborative innovation among HEIs	1. Identification of ICT tools to use 2. Training 3. Curriculum aligning use of ICT tools	1. Change the curriculum to meet the usage of ICT tools 2. Innovative curriculum 3. Collaborations of the HEIs in innovation projects

From the interviews, it was evident that teachers' concern related to infrastructure improvement. Curriculum change and innovative curriculum emerged new themes aimed to suit the introduction of ICT tools.

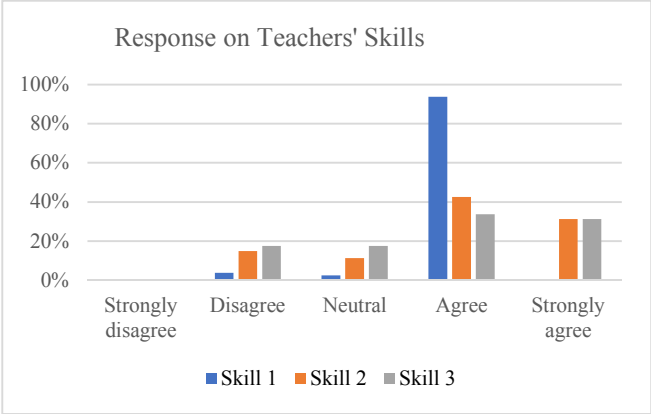
C. Initial Stage Factors- Subject

The results regarding the adequacy of ICT infrastructure were checked against construct as in figure 3 as follows:



Items	Statements
Item 1	The local area network (LAN) is available and works well
Item 2	The Internet and services are adequately working well
Item 3	The bandwidth provide is enough and serves the users well
Item 4	The wireless access is abundantly available for teachers and students

Fig. 3. Adequacy of ICT infrastructure

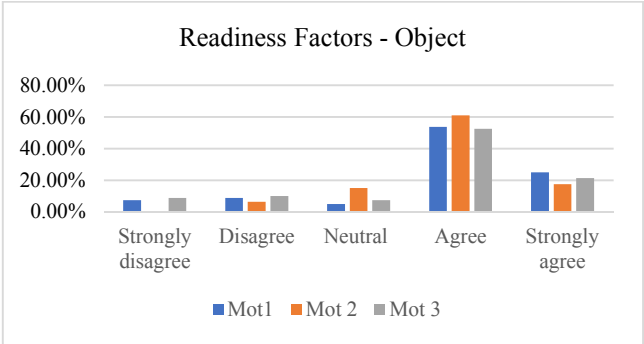


Skills	Statements
Skill 1	I have enough skills in using computers, laptops, and mobile devices
Skill 2	I am able to find, download, share, and use educational materials for my students
Skill 3	I have the skill to find and use different ICT tools in classrooms

Fig. 4. Teachers' skills responses

The teachers showed skills to the number of constructs set for this study. Especially, they showed to possess skills in using computers, laptops, and mobile devices.

D. Readiness Factors - Object



Motivation	Statements
Mot1	It is a good idea to change teaching style to an innovative one
Mot2	The benefits of using ICT tools are and should be applied in teaching and learning properly
Mot3	I am ready to use ICT tools in classrooms

Fig. 5. Responses on the readiness factors

The teachers almost all of them have responded to show readiness to integrate ICT tools in their teaching. What remains is the management of HEIs to provide the conducive environment for the teachers to work.

V. DISCUSSION

Three research questions (RQs) were asked to fulfill the objectives of the study. RQ1. Asked to what extent is the current ICT infrastructure and resources adequate to enable ICT integration in the teaching and learning in in higher education institutions in Tanzania? The responses of the teaching staff and IT staff showed generally to agree on the constructs related to the initial stage factors. The items 2 and 4 though indicated somehow not well scores compared to items 1 and 3. Almost 40% disagreed on item 4 suggesting that the wireless access is not abundantly available for teachers and students. RQ2 asked what are the ICT skills and readiness of teachers of higher education institutions in Tanzania in integrating ICT tools for improving innovative teaching and learning? The responses showed mostly teachers have enough skills in using computers, laptops, and mobile devices. The skills to find, share, download, and use ICT tools were good almost for all teachers. The interviews were related to the RQ3 which asked what are the teachers' views on the different ways for the improvement of innovative teaching and learning in higher education institutions in Tanzania? The responses indicated through themes produced that training and improvement of the ICT infrastructure are of crucial importance. Furthermore, a change in curriculum to really reflect innovation should be established, and the final theme of collaboration in HEIs innovations agreed as part of the successful integration of ICT tools in HEIs in Tanzania.

VI. CONCLUSION AND FUTURE WORK

Our study has highlighted important issues to consider when improving innovative teaching and learning in HEIs in Tanzania. At this moment when technologies, creativity, and innovations are crucial issues for the success of any society, HEIs should lead and show the way. The consideration of initial factors and readiness factors shown in our research model have provided a base for the policymakers in HEIs and education sector generally to the improvement of innovative teaching and learning. The future work is expected to develop the curriculum which really provides a chance for students to be innovative. The type of teaching and learning that enrich teachers and students to be innovative. Furthermore, to conduct studies that directly solves societal challenges through innovations.

ACKNOWLEDGMENT

We would like to thank all the participants from the three HEIs for their contributions towards the completion of this paper. The information they provided is pivotal to innovation mission in HEIs in Tanzania. We also would like to thank the management of the three institutions in Tanzania for agreeing to let their staff participate in the study.

REFERENCES

[1] P. Kihoza, I. Zlotnikova, J. Bada and K. Kalegele, "Classroom ICT integration in Tanzania: Opportunities and challenges from the perspectives of TPACK and SAMR models," *International Journal of Education and Development Using Information and Communication Technology,* vol. 12, no. 1, pp. 107-128, 2016.

[2] J. S. Mtebe, H. Dachi and C. Raphael, "Integrating ICT into teaching and laerning at the University of Dar es Salaam," *Distance Education,* vol. 32, no. 2, pp. 289-294, 2011.

[3] X. Muainga, H. Hansson, A. Nilsson, A. Mondlane, I. Mutimucio and A. Guambe, "ICT in Education in Africa - Myth or reality: A case study of Mozambican Higher Education Institutions," *The African Journal of Information Systems,* vol. 5, no. 3, pp. 106-117, 2013.

[4] M. P. Mahenge and C. Sanga, "ICT for e-learning in three higher education institutions in Tanzania," *Knowledge Management & E-Learning,* vol. 8, no. 1, pp. 200-212, 2016.

[5] F. W. K. Amenyedzi, M. N. Lartey and B. M. Dzomeku, "The use of computers and Internet as supplementary source of educational materials: A Case study of senoir high schools in the Tema Metropolis in Ghana," *Contemporary Educational Technology,* vol. 2, no. 2, pp. 151-162, 2011.

[6] The United Republic of Tanzania, "National Information and Communications Technology Policy," Ministry of Works, Transport and Communication, Dar es Salaam., 2016.

[7] Ministry of Education and Vocational Training, "Education and Training Policy," United Republic of Tanzania Ministry of Education and Vocational Training, Dar es Salaam, 2014.

[8] United Republic of Tanzania, "Tanzania Development Vision 2025," Ministry of Finance and Planning, Dar es Salaam, 2000.

[9] K. Akyeampong, K. Lussier, J. Pryor and J. Westbrook, "Improving teaching and learning of basic school mathematics and reading in Africa: Does teacher preparation count?," *International Journal of Educational Development,* vol. 33, no. 3, pp. 272-282, 2013.

[10] S. Sarkar, "The role of Information Communications Technology (ICT) in higher education for the 21st Century," *The Science Probe,* vol. 1, no. 1, pp. 30-41, 2011.

[11] L. Johnson, B. S. Adams, M. Cummins, V. Estrada, A. Freeman and C. Hall, "NMC Report: 2016 higher education edition," The New Media Consortium, Austin, Texas, 2016.

[12] C. McLoughlin and M. J. W. Lee, "Personalised and self regulated learning in the Web 2.0 era: International examplars of innovative pedagogy using social software," *Australasian Journal of Educational Technology,* vol. 26, no. 1, pp. 28-43, 2010.

[13] Y.-J. Lee, "A study on the effect of teaching innovation on learning effectiveness with learning satisfaction as a mediator," *World Transactions on Engineering and Technology Education,* vol. 9, no. 2, pp. 92-101, 2011.

[14] G. Sang, M. Valcke, J. v. Braak and J. Tondeur, "Student teachers' thinking processes and ICT integration: Predictors of prospective teaching behaviours with educational technology," *Computers and Education,* vol. 54, pp. 103-112, 2010.

[15] M. Tedre, F. Ngumbuke and J. Kemppainen, "Infrastructure, human capacity, and high hopes: A decade of development of e-learning in a Tanzanian HEI," *Redefining the Digital Divide in Higher Education,* vol. 7, no. 1, pp. 1-15, 2010.

[16] UNESCO, "ICT competence framework for teachers," United Nations Educational Scientific and Cultural Organization, Paris, 2011.

[17] K. Kuutti, "Activity theory as a potential framework for human-computer interaction research," in *Context and consciousness: Activity theory and human-computer interaction*, Cambridge, MA, MIT Press, 1996, pp. 9-22.

[18] N. H. Hashim and M. L. Jones, "Activity theory: A framework for qualitative analysis," in *4th International Qualitative Research Convention (QRC)*, PJ Hilton, Malaysia, 2007.

[19] W. P. Mtega, R. Bernard, A. C. Msungu and R. Sanare, "Using mobile phones for teaching and learning purposes in higher education: A case of the Sokoine University," in *The 5th UbuntuNet Alliance Annual Conference*, Dar es Salaam, 2012.

[20] A. Kafyulilo, P. Fisser and J. Voogt, "ICT use in science and mathematics teacher education in Tanzania: Developing technological pedagogical content knowledge," *Australasian Journal of Educational Technology,* vol. 31, no. 4, pp. 381-399, 2015.

[21] C. Kivunja, "Innovative methodologies for 21st Century learning, teaching and assessment: A convenience sampling investigation into the use of social media technologies in higher education," *International Journal of Higher Education,* vol. 4, no. 2, pp. 1-26, 2015.

[22] G. Conole and P. Alevizou, "A literature review of the use of Web 2.0 tools in higher education," The Higher Education Academy, Milton Keynes, UK., 2010.

[23] G. I. Mwandosya and C. Suero Montero, "Towards a Mobile Education Tool for Higher Education Teachers: A User Requirements Definition," in *Proceedings of the 2017 IEEE Science Technology & Innovation Africa Conference*, Cape Town, 2017.

[24] M. Ford and T. Leinonen, "MobilED- A mobile tools and services platform for formal and informal learning," in *mLearn 2006, the 5th World Conference on Mobile Learning*, Banff, Canada., 2006.

[25] G. Fauville, A. Lantz-Andersson and R. Säljö, "ICT tools in environmental education: reviewing two newcomers to school," *Environmental Education Research,* pp. 1-36, 2013.

[26] E. T. Lwoga, "Making learning and web 2.0 technologies work for higher learning institutions in Africa," *Campus-Wide Information Systems,* vol. 29, no. 2, pp. 90-107, 2012.

[27] P. A. Ertmer and A. T. Ottenbreit-Leftwich, "Teacher technology change: How knowledge, confidence, beliefs, and culture intersect," *Journal of Research on Technology in Education ,* vol. 42, no. 3, pp. 255-284, 2010.

[28] B. Makki and B. Makki, "The impact of integration of instructional systems technology into research and educational technology," *Creative Education,* vol. 3, no. 2, pp. 275-280, 2012.

[29] A. Mwalongo, "Teachers' perceptions about ICT for teaching, professional development, administration and personal use," *International Journal of Education and Development Using Information and Communication Technology,* vol. 7, no. 3, pp. 36-49, 2011.

Smart Energy Generation for a Smart Campus

Esrom Mahlatsi Malatji
Dept. of Electrical and Electronic Engineering Technology
University of Johannesburg
Johannesburg, South Africa
emalatji@uj.ac.za

Sula Bantubakhona Kwesi Ntsaluba
Dept. of Electrical and Electronic Engineering Technology
University of Johannesburg
Johannesburg, South Africa
sulan@uj.ac.za

***Abstract*—As cities move closer to the concept of smart, the idea of a smart campus came into being. One of the most important characteristics of a smart campus is the way in which the campus interacts with the environment. Clean energy contributes to a clean environment and the research in this paper looks at ways to supply a campus with clean energy from renewable energy sources. This is referred to as smart energy generation due to its positive impact on the environment. One of the challenges with the usage of renewable sources which is addressed in this paper is the difficulty in supplying an entire load due to its stochastic nature. It was found out through using a combination of renewable energy sources and energy storage, a whole campus can be supplied with clean energy without additional energy being drawn from the grid.**

Keywords—battery storage, biogas, renewable energy, solar PV, smart campus

I. INTRODUCTION

As the world prepares itself for the fourth industrial revolution which will reshape the way we live, the concept of smart cities arises as one integral part of this revolution. From the concept of smart cities, a smart campus concept has been developed, which take into account performance indicators and characteristics of smart campus which are shown in Fig. 1 [1-4].

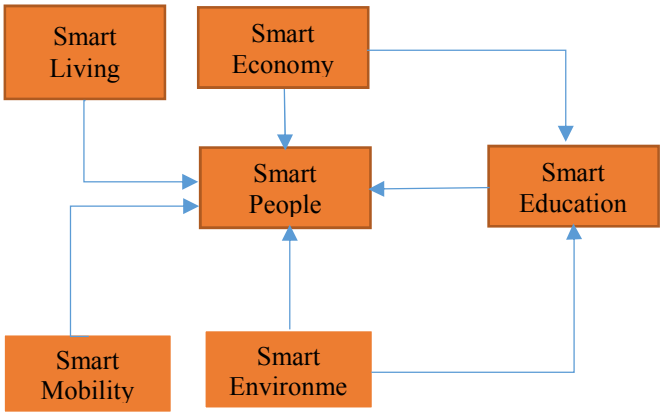


Fig. 1. Structural model of a smart campus

The characteristic that is relevant to this research is the smart environment. The smart environment is about the way the people interact with the environment. The way the campus uses and manages its resources such as water, waste and energy. The campuses should implement waste recycling initiatives which will contribute positively to the environment. Integration of Renewable energy sources to smart campuses if they seek to be self-sufficient with regard to energy production. One such campus that is almost self-sufficient when it comes to energy production is the University of California campus in San Diego. This campus produces almost 100% of its energy [5]-[7]. The other complex similar to a campus that has implemented energy systems in order to become self-sufficient when coming to energy requirements is the military facilities around the world [8]-[10]. Most of these facilities make use of Solar PV, heat recovery and battery energy storage systems (BESS). The research in this paper will focus on designing a smart energy system for a smart campus. The system will consist of a combination of solar PV, biogas and battery storage. Such systems were designed for different applications and most of these research used solar PV, wind, batteries and diesel generator [11].

In this case, a diesel generator is being utilized as a supply for a base load as the other resources are stochastic in nature. The disadvantage of using a diesel generator lies in its environment unfriendliness. Further to this, in most parts of the world diesel can be a very expensive type of fuel. The research in this paper will seek to utilise biogas as an environmentally friendly alternative to diesel generators because biogas can be utilised to supply base load. Biogas can be generated from organic material like sewage, abattoir waste and food waste. The research in this paper only focuses on the food waste as it is readily available in metropolitan areas. Food waste can be collected from the campuses and places like restaurants and shopping malls that are nearby the campus. Most campuses are situated in areas that have a potential to generate food waste.

In this paper a comprehensive smart energy generation is presented which is in line with requirements of smart campus. Unlike most smart campuses, this campus will not depend on the grid in any way. Energy will be supplied from the following sources: solar, biogas and the battery storage. The sizing of a biogas plant will reveal how much gas is needed to produce a required amount of power and this is very rare in existing literature.

The rest of the paper is organized as follows: the development of a smart energy generation for a smart campus is presented in section II, a case study is then presented in section III. A different design of a different section of the smart generation is presented in section IV. The results are presented and discussed in section V Finally, concluding remarks are presented in section VI.

II. DEVELOPING A SMART ENERGY GENERATION FROM A SMART CAMPUS

One of the characteristics of a smart energy system is the smart environment, which requires a campus to have a majority or all its energy generated from renewable sources. Most of the power consumption within a campus happens

during the day, which makes solar energy one of the preferred choices. The challenge presented with the solar energy option is availability of energy outside periods where sunlight is available, resulting in a need to couple this renewable energy source with another. Biogas can be used as on other option to supply the base load energy required. The campus selected for the case study is that of the University of Johannesburg, South Africa. The location of this campus makes biogas a feasible option because of the high volumes of food waste that is generated within the city. It worth mentioning that biogas produces a lot of noise. Therefore, an innovative way to suppress this noise must be devised. The remaining load can be catered for by making use of the energy storage in the form of batteries. Battery storage is a feasible option for the environment of the campus. Fig.2 shows the smart generation with all the DC generation being converted to AC as the campus load is an AC load.

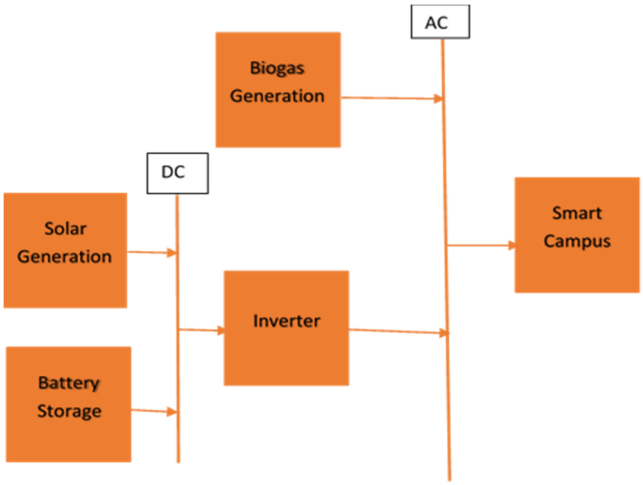


Fig. 2. Smart energy generation

III. Case Study

The university of Johannesburg (UJ) will be used as a case study in this research project. The university has multiple campuses across Johannesburg, with a campus at Soweto, two campuses at Auckland park and one at Doornfontein in the heart of Johannesburg. The Auckland park campus is being used in this case study because is considered as the main campus. This campus is energy intensive especially during the day as most activities happen during the day. Fig.3 shows the load of this campus.

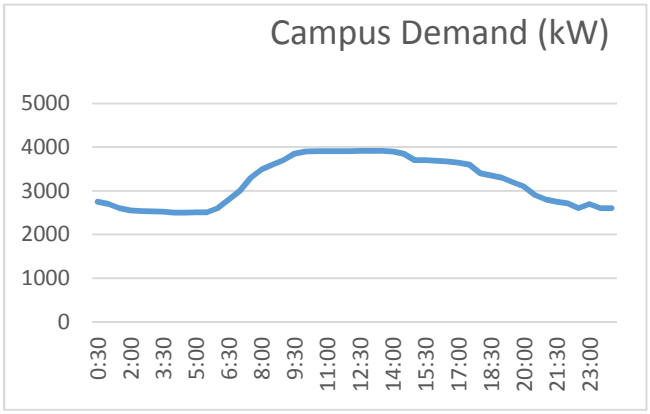


Fig. 3. The load of the campus

IV. Designing the Power Generation for a Smart Campus

Three different supplies are used for smart energy generation. These sources are solar photovoltaic, biogas generation from mainly food waste, and battery storage. The technology type for this source is not of importance for this research. Another aspect that was not of interest for this research is whether the biogas plant is continuous or batch, as the main objective is to show the feasibility of implementing the biogas plant.

A. Solar Photovoltaic(PV)

The photovoltaic generator generates power according to eq. 1, which is the power generated at time t

$$P_{solar}(t) = \eta_{panel} A_{total} I_{irradiance}(t) \quad (1)$$

The total surface area of the solar panel is denoted by A_{total} and is given in eq.2.

$$A_{total} = N_{total}(L_{panel} \times W_{panel}) \quad (2)$$

Where N_{total}, L_{panel} and W_{panel} are is the total number of panels, the length and the width of the solar panel respectively. The efficiency of the panel is calculated from eq.3 and is adapted from [12].

$$\eta_{panel} = \eta_{GE}\left[1 - 0.9\alpha\left(\frac{I_{irradiance}}{I_{irradiance,norm}}\right)(T_{c,norm} - T_{A,norm}) - \alpha(T_A - T_R)\right] \quad (3)$$

From eq. 3, irradiance is denoted as $I_{irradiance}(t)$. The PV generator efficiency at reference temperature T_R (25^0 C) is denoted as η_{GE} is what is refered as the photovoltaic generator efficiency, The temperature efficiency is given by α which has a value ranging from 0.004-0.006^{0}C.The average irradiance from the sun is given by $I_{irradiance}$ and it has a value in the region of 0.8kW h/m^2 at a nominal (norm) temperature. The other two temperatures $T_{A,norm}$ and $T_{c,norm}$ are the ambient and cell respectively. Fig. 4 shows the power generated by solar while figure 5 show the solar generation together with the campus load which clearly highlights the regions where solar would not be able to supply. These regions are between 18:30 in the evening and 06:30 in the morning where the solar output is zero while the load is about 2000 kW.

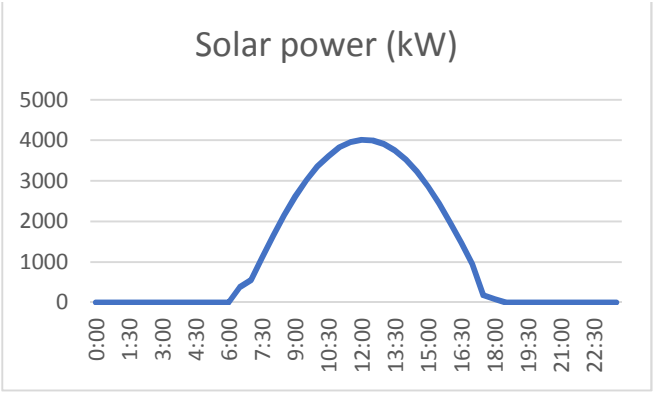


Fig. 4. Solar generation

B. Biogas

As mentioned most campuses generate a sizeable quantity of food waste. Most campuses are located in the cities and there are a lot of activities which generate a lot of food waste. This is a good source for feed to a biogas plant. Biogas will serve as a base load for the campus. Biogas is an organic gas that is made mostly of methane and carbon dioxide. Methane is combustible gas which can be used as fuel which can be used in a gas generator to generate electricity. Because most campuses are within city boundaries, building a biogas plant can have environmental challenges. An innovate design must be considered to counteract these challenges. The biogas plant is mostly dependent on the quality of the feed to the plant and the mechanics of the biogas engine. The output power of the biogas plant is given by the eq.4. Fig.6 shows the load deficiency which is denoted as the load that the solar plant cannot supply. This load must be supplied by the biogas plant.

$$P_{bg}^{min} \leq P_{bg}(t) \leq P_{bg}^{max} \quad (4)$$

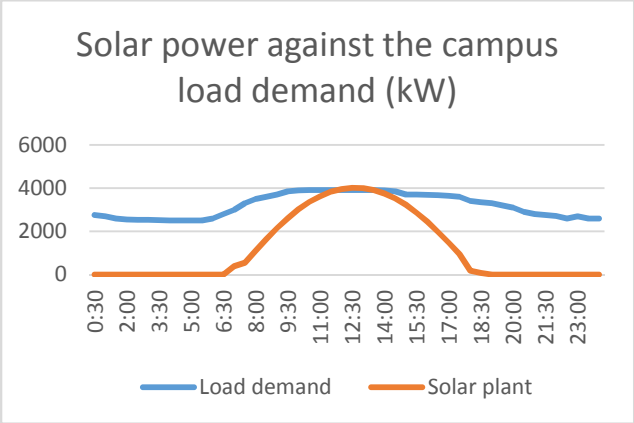


Fig. 5. Solar generation with campus load

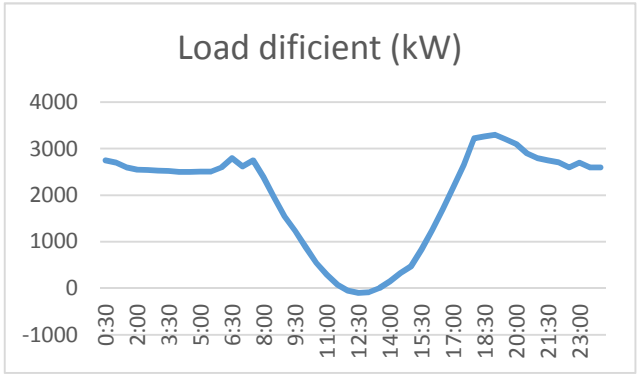


Fig. 6. Load that cannot be supplied by solar

At the heart of a biogas plant is the biogas engine which needs biogas feed. The Biogas gas plant must generate 3MW of power, which must be generated from food waste. The first part is to calculate the amount of food waste that can generate enough gas that will be able to produce 3MW. To find the amount of food waste needed depends on the specifications of the engine, the engine considered is Jenbacher gas engine JMS 320 GS-B.L. This engine has a history dating back to 1988 and is among the best in the biogas industry.

Table 1 gives the specifications of the engine [13]. Just like in internal combustion engines, most of the energy in the biogas engine is converted into heat and some of this heat is indicated as heat to be dissipated in Table 1. The fraction of the energy that does work is known as the thermal energy. The thermal efficiency can be considered as the performance measure of the biogas engine. The energy input is the mechanical power of the engine and the efficiency of the engine is 40.8%. The energy output of the engine is the electrical power which is 1.063 KW as given in Table 1. The Fuel consumption based on Low Heating Value (LHV) refers to the amount of fuel that needed in the engine in order to produce the rated power output of the engine.

TABLE I. ENGINE SPECIFICATIONS

JMS 320 GS-B.L Specification	
Heat to be dissipated	67 kW
Thermal efficiency	42.3%
Total efficiency	83.1%
Electrical efficiency	40.8%
Energy Input	2.606 kW
Electrical output	1.063 kW el.
Recoverable thermal output (180°)	1.103 kW
Fuel consumption based on a LHV	4.5 kWh/Nm^3

$$E_{fuel} = \frac{E_{electrical}}{\eta} = \frac{1063\ kW}{0.408} = 2605\ kW \quad (5)$$

$$\dot{m} = \frac{E_{fuel}}{LHV} = \frac{2605kW}{4.5kWh/Nm^3} = 579Nm^3/h \quad (6)$$

To produce 1 MW of power, *579 Nm^3* of biogas per hour is needed which amounts to 13893 *Nm^3* if the plant is to run for 24 hours in a day. The research conducted in [14]-[15] , gives the yield of biogas in litres per kg of various materials. The food waste yields 160 litre/kg and to produce 13893 *Nm^3* of gas approximately 90 000kg of food is needed. The biogas plant does not have to run for 24 hours because some hours will be supplied by the solar plant. Figure 7 shows the load that has to be supplied by the biogas plant. The biogas plant needs to supply a load in the region of 2.6 MW from 08h30 in the evening to 06h30 in the morning and it needs to operate at 3 MW from 17:30 to 20:30 as this is the peak period.

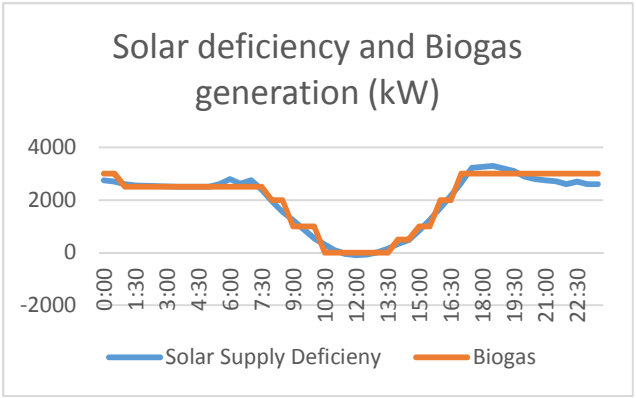


Fig. 7. Biogas supplying solar deficiency

C. *Battery Storage*

One of the major challenges with renewable energy sources is that there are stochastic in nature, resulting in times were the sources are not available or are not in adequate for the load. The base load will be supplied by the biogas plant and solar plant we supply the load during the day. There are times where the load that cannot be supplied by both the solar and biogas plants, more notably during the peak periods which happen between 17:30 and 20:30. This demand can be catered for by making use of battery storage. Battery storage is one of the most commonly used storage as of recent years. There is numerous research and improvements in the area of battery storage. During low peak periods, excess power can be stored in the batteries which can be used during high peak periods. The storage needs to be able to supply energy of 1250 kWh a day. To conserve the life span of the batteries, batteries should not discharge 100% of their charge as they should discharge to between 40 to 50% of their storage capacity. The size of the battery storage that should be used in this research is the 2500kWh battery storage which is double the size of the energy required. The battery charges according to eq.7 which is obtained from [16-17]. Eq.8 denotes the constraints of the battery, while the depth of charge is denoted by eq.9. The load that is supposed to be supplied by the battery storage is shown in fig. 8.

$$B_c(t) = B_c(t-1) + \eta_c P_1(t) - \eta_D P_2(t) \quad (7)$$

$$B_c^{min} \le B_c(t) \le B_c^{max} \quad (8)$$

$$DOD = 1 - \frac{B_c^{min}}{B_c^{max}} \quad (9)$$

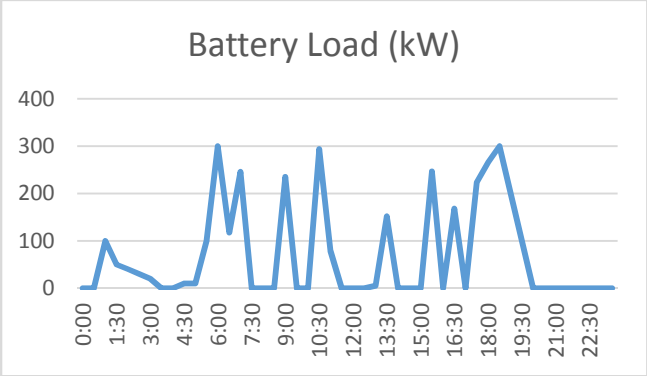


Fig. 8. Load to be supplied by the storage

Figure 8 give the load that would be supplied by the battery storage.

V. Results and Discussions

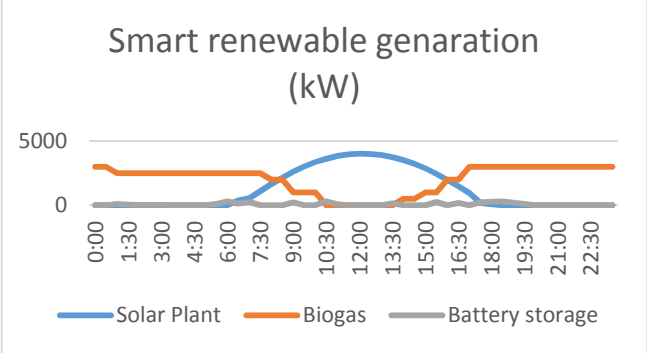


Fig. 9. The renewable energy generation

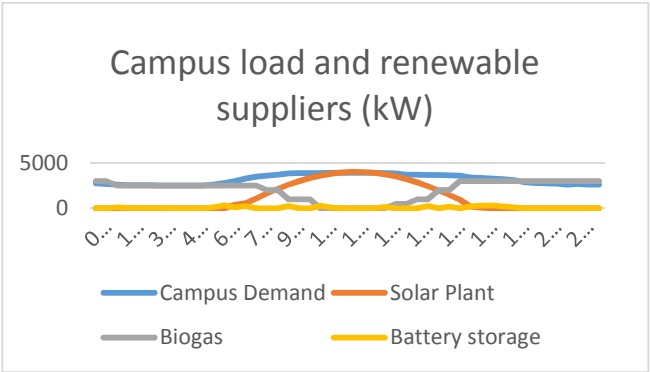


Fig. 10. The renewable energy generation together with the camus load

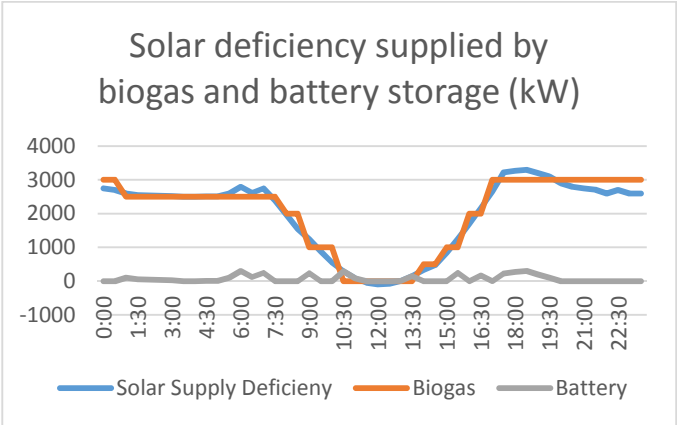


Fig. 11. Biogas and storage suppling the load that can not be supplied by soalr.

Fig. 9 gives the generation from solar , biogas and the battery storage with solar energy being significate during the day. This is due to the fact that solar resources are available during this period. The biogas provides the base load which is evident during the night. The battery storage provides energy in those rare cases when both solar and biogas are unable to supply. The contribution of the storage is very minimal as seen from the fig. 9.

Fig.10 shows the campus load with the contributions of different renewable energy supply. The contribution of the different energy sources to the load is clearly visible from this figure The time during midday where the solar energy is in suppliers tht s the energy that it is used to charge the battery for storage. Fig. 11 shows the performance of the biogas and the battery storage with the load deficient that can be supplied by the solar plant.

VI. Conclusions

This paper presented a smart energy generation concept designed for a smart campus. The renewable energy sources that were used include: solar, biogas and battery storage. The system supplied energy to a campus with a 4 MW maximum demand. The solar energy supplied energy during the day while the biogas supplied energy mainly at. The results of the study illustrated that it is possible to completely supply a campus with off-grid energy provided that the system is optimized taking into account the different energy sources and their corresponding constraints.

REFERENCES

[1] W. Muhamad, B. K. Novianto, and S. Yazid. "Smart campus features, technologies, and applications: A systematic literature review." *International Conference Information Technology Systems and Innovation (ICITSI)*, Bandun Jawa Barat, 2017

[2] E. Malatji, "The development of a smart campus-African universities point of view." 8th International Renewable Energy Congress (IREC), 21-23 Mar. 2017 Amman, Jordan [Online]. Available: IEEE Xplore

[3] I. S. Saina, "My AppCorner: personal interaction areas on public multi-touch displays in a smart campus environment". MS [Dissertation]. Tampere, Finland : Tampere Univ.2018. [Online]. Available: Tampere university institutional repository

[4] V. B. Ribeiro, et al. "Integration and intelligence for sustainability—I2S An overview of the I2S: Intelligent building's concept from the Campos dos Goytacazes innovation hub campus." *URUCON*, 23-25 Oct. 2017,Montevideo, Urugay [Online]. Available: IEEE Xplore

[5] *B. Washom* , J Dilliot, A. Weil, J. kleissl, N. Balac and C. Richter., "Ivory Tower of Power: Microgrid Implementation at the University of California, San Diego," in *IEEE Power and Energy Magazine*, vol. 11, no. 4, pp. 28-32, July-Aug. 2013.

[6] M. Lave, J. Kleissl, and J. Stein, "A wavelet-based variability model (WVM) for solar PV powerplants," *IEEE Trans. Sustainable Energy*, vol. 4, pp. 501 -509 , Aug. 2012.

[7] C. W. Chow, B. Urquhart, J. Kleissl, M. Lave, A. Dominguez, J. Shields, and B. Washom, "Intra-hour forecasting with a total sky imager at the UC San Diego solar energy testbed," *Solar Energy*, vol. 85, pp. 2881-2893, Nov., 2011.

[8] S. B. Van Broekhoven N. Judson S. V. T. Nguyen and W. D. Ross "Microgrid study: Energy security for DoD installations " MIT Lincoln Lab. Lexington MA Tech. Rep. TR 1164 2012.

[9] White House Press Release. Fact Sheet: Obama administration announces additional steps to increase energy security, September 10, 2018 [Online]. Available: http://www.whitehouse.gov/the-press-office/2012/04/11/fact-sheet-obama-administration-announces-additional-steps-increase-energy [Accessed September 10, 2018].

[10] S. Van Broekhoven, N. Judson, J. Galvin and J. Marqusee, "Leading the Charge: Microgrids for Domestic Military Installations," in *IEEE Power and Energy Magazine*, vol. 11, no. 4, pp. 40-45, July-Aug. 2013.

[11] H. Tazvinga, Z. Bing, and X Xia. "Energy dispatch strategy for a photovoltaic–wind–diesel–battery hybrid power system." *Solar Energy* vol. 108, no. 1, pp. 412-420, Aug. 2014.

[12] Z. Wu, H. Tazvinga, and X. Xia. "Demand side management of photovoltaic-battery hybrid system." *Applied Energy*, vol. 148, no. 1 , pp. 294-304, Aug. 2014.

[13] General Electric, Technical Describtion, September 10, 2018. [Online]: Available: http://www.provincia.livorno.it/new/spawdocs/ambiente/Technical%20Description_AB%20Energy%20320.pdf . [Accessed September 10, 2018].

[14] C. Popescu and S. Jurcoane, "Evaluation of biogas potential of some organic substrates from agriculture and food industry and co-digestion in large scale biogas plant*" Romanian Biotechnological Letter*, vol. 20, no. 4, pp-10648-10655, Nov. 2015.

[15] R. Kigozi, A. Aboyade and E. Muzenda, "Biogas production using the organic fraction fraction of municipal solid waste as feedstock" *International journal of research in chemical, metallurgical and civil engineering*, vol. 1, no.1, pp. 107-114, Jun. 2014.

[16] H. Tazvinga, X. Xia and J. Zhang, "Minimum cost solution of photovoltaic-diesel-battery hybrid power systems for remote consumers" *Solar Energy*, vol 96, no. 1, pp. 292-299, Aug. 2013.

[17] H. Tazvinga, Z. Bing, and X Xia. "Optimal power flow management for distributed energy resources with batteries." *Energy Conversion and Management*, vol. 102, no. 1, pp. 104-110, Sep. 2014.

Modeling the Influence of Large-Scale Circulation Patterns on Precipitation in Mauritius

Kreshna Gopal
Scinova Consulting Ltd.
Mauritius
kgopal@scinova.com

C Prakash Khedun
Scinova Consulting Ltd.
Mauritius
pkhedun@scinova.com

Anoop Sohun
Scinova Consulting Ltd.
Mauritius
asohun@scinova.com

***Abstract*—Mauritius suffers from chronic water shortages that can severely impact its economy and the well-being of its population. Both surface and groundwater availability are determined by rainfall, which is in turn influenced by large-scale circulation patterns such as the El Niño Southern Oscillation (ENSO) and the Indian Ocean Dipole (IOD). Here we report on the influence of these two teleconnection patterns and present the result of a simple neural network for precipitation forecasting, based on the state of ENSO and IOD. Data from the Vacaos station, for the period 1961 to 2012 is used. We found statistically significant correlation between average winter rainfall and ENSO and IOD indices. The correlation for summer was negligible. The prediction of summer precipitation was less accurate than that of winter precipitation. The findings from this study can help in more efficient planning and management of water resources on the island.**

Keywords—precipitation, ENSO, IOD, neural network.

I. Introduction

The water shortages that Mauritius regularly faces is alarming. Based on the United Nations' definition [1], Mauritius is a water stressed country. The Water Resources Unit in Mauritius predicts that by 2020, the country will slip into the water scarce category [2]. Furthermore, the island does not have enough carry-over capacity, i.e. most of the water received during a hydrologic year is used within the year itself. Thus, a shortage of rainfall can have immediate and disastrous effects on the well-being of the population.

We conducted a comprehensive study [3] where we (i) investigated the relationship between the El Niño Southern Oscillation (ENSO) and the Indian Ocean Dipole (IOD) with precipitation, (ii) developed a precipitation forecasting model using artificial neural network, and (iii) conducted a bivariate drought analysis using multiple precipitation deficit variables (duration, severity, and inter-arrival time).

In this paper, we present one component of this study, which is establish the relationship between large-scale circulation patterns (ENSO and IOD) on precipitation in Mauritius and use these indices to predict precipitation using a neural network. A predictive model can enable medium- and long-term planning and facilitate proactive measures for forecasted water shortages and droughts.

II. Data

A. Precipitation Data

Monthly precipitation data for the Vacoas station was obtained from two sources. Data for the period from 1961 to 1990 was obtained from the National Climate Data Center (NCDC) which is part of the Global Historical Climatology Network (GHCN). Data beyond 1990 was not available from the NCDC and was purchased from the Mauritius Meteorological Services (MMS). Fig. 1 gives a plot of the monthly precipitation for January 1961 to September 2012, along with the long-term mean (170.54 mm). It is apparent that the precipitation is not uniform or close to the long-term mean but is highly variable. For most months, the precipitation is below the long-term mean, and for some months, the precipitation is several magnitudes above the long-term mean.

The largest variation in precipitation is noted for February followed by January, which are also the two wettest months. One notable outlier is in December, where in December 1961, the monthly total was 1,362 mm. The median for precipitation is generally lower than the long-term mean for each month, which implies a slight positive skewness in monthly precipitation.

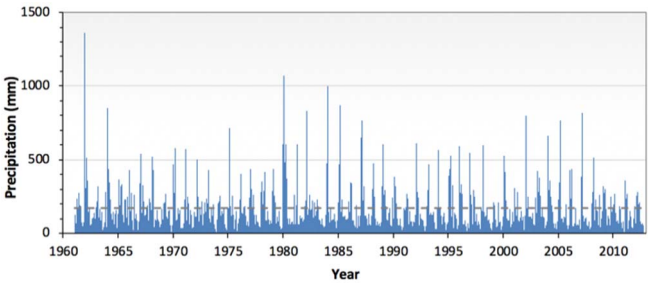


Fig. 1. Monthly precipitation at Vacoas (1961 – 1990). The dotted gray line represents the long-term mean (170.54 mm) for the dataset.

B. El Niño Southern Oscillation

Several indices are available for ENSO, including the Southern Oscillation Index (SOI), Niño 3, Niño 4, Niño 3.4, the Multi ENSO Index (MEI), among others. The Niño 3.4 index [4] has been employed in a number of hydro-meteorological studies because it often exhibits the strongest influence and correlation on remote hydro-climatological events. It is the area-averaged sea surface temperature anomaly (SSTA) over the region bounded by 5°N–5°S and 120°W–170°W.

Trenberth [5] also proposed the use of the Niño 3.4 index for the definition of El Niños, and an "El Niño can be said to occur if 5-month running means of sea surface temperature (SST) anomalies in the Niño 3.4 region exceed 0.4°C for 6 months or more." The opposite can be adopted for the definition of La Niña events. Other definitions, with slight variations of the one proposed in [5], have since been proposed and adopted in different regions of the world. No official definition for El Niño or La Niña has been developed or adopted for Mauritius.

Monthly data for the Niño 3.4 index was obtained from the International Research Institute on Climate and Society Data Library. Fig 2 shows a plot of the Niño 3.4 index for the period January 1961 to July 2009. Sustained positive (negative) anomalies are indicative of El Niño (La Niña) conditions. The Niño 3.4 index is overlain with a 5-month moving average and the ±0.4 thresholds.

Funding for this study was provided by the Mauritius Research Council (MRC) under the African Adaptation Programme [MRC/RUN/AAP-1210]

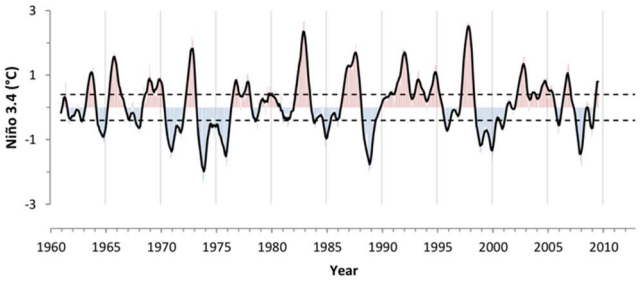


Fig. 2. Time series of Niño 3.4 index smoothed with a 5-month moving average and ±0.4 thresholds (broken lines).

C. *Southern Oscillation Index*

The Southern Oscillation Index (SOI) is an index that shows the variation in the Southern Oscillation. It is the difference in mean sea level pressure anomalies at Tahiti and Darwin, Australia. Data for the SOI was obtained from the National Center for Atmospheric Research (NCAR) Climate and Global Dynamics (CGD).

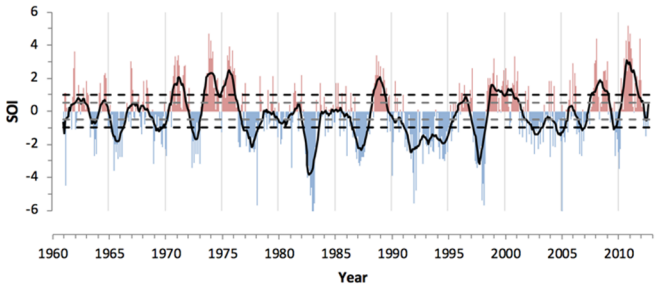


Fig. 3. Time series of the Southern Oscillation Index (SOI) overlain with a 13-month moving average and ± 0.5 and ± 1 thresholds.

Fig. 3 shows a plot of SOI for the period 1961 to 2012. The data is smoothed with a 13-month centered moving average to highlight the positive and negative phases. Note that the SOI is negatively correlated with the Niño 3.4 index, i.e. an El Niño (La Niña) will be represented by positive (negative) Niño 3.4 indices and the SOI will be negative (positive). The strength of El Niño and La Niña events can be gauged based on the scale given in Table I.

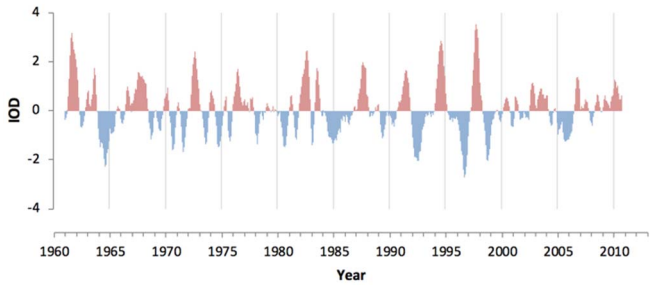


Fig. 4. Time series for IOD.

D. *Indian Ocean Dipole*

IOD is defined as the anomalous SST gradient, known as the Dipole Mode Index (DMI), between the western equatorial Indian Ocean (50°E-70°E and 10°S-10°N) and the south eastern equatorial Indian Ocean (90°E-110°E and 10°S-0°N). DMI data was obtained from the Japan Agency for Marine-Earth Science and Technology (JAMSTEC). Fig. 4 is a plot of the DMI for the period January 1961 to September 2010. Data beyond the latter date was not available.

III. Relationship Between Large-Scale Circulation Patterns and Precipitation

The correlation between the climate indices and precipitation was determined using Pearson correlation. The correlation coefficient, ρ_{xy}, is a measure of linear association between two time series x and y and is given by

$$\rho_{xy}(k) = \frac{E\left[(x_i - \mu_x)(y_{i+k} - \mu_y)\right]}{\sigma_x \sigma_y} \qquad (1)$$

where $\rho_{xy}(k)$ is the cross-correlation for lag k between time series x_i and y_i with means μ_x and μ_y and standard deviation σ_x and σ_y respectively, and $E[\cdot]$ is the expectation operator. The range for $\rho_{xy}(k)$ is [-1, 1], with larger $|\rho_{xy}|$ implying stronger association between *x* and *y*.

Mauritius has two seasons: winter and summer. Winter lasts from May to October and summer is from November to April. November is also the beginning of the hydrological year. In order to assess the influence of ENSO and IOD on precipitation at Vacoas, the precipitation series was divided into two series, winter and summer, and the average precipitation for each year was computed. The Niño 3.4, SOI and IOD series were similarly divided into two sets.

The correlation between Niño 3.4, SOI, and IOD with average winter and summer precipitation were computed. The magnitude and sign of the correlation coefficient indicates the existence, strength, and nature of any influence the teleconnection phenomena have on precipitation.

A. *Correlation between Precipitation and SOI*

Table I gives the correlation coefficient and *p*-values for the correlation between precipitation anomaly and the two ENSO indices, which is the difference between the average precipitation for each year and the long term mean for the season. We divided the data into two periods (1961-1990 and 1991-2009) because they come from different sources. We also do the analysis for the combined periods (1961-2009). No trend is apparent in the summer plot, but a small negative trend, above average rainfall with negative SOI, is visible in the winter plot.

Note that the correlation between Niño 3.4 and precipitation is of opposite sign compared to SOI and precipitation, because Niño 3.4 and SOI are negatively correlated.

There exists a small correlation between winter indices and precipitation, for the 1961-1990 series, while there is no apparent correlation between the indices and summer precipitation. The correlation is not statistically significant at an alpha of 0.05 but statistically significant at an alpha of 0.1. The correlation between SOI and winter precipitation is stronger than that between Niño 3.4 and precipitation. Even though both phenomena are related, stronger correlation with SOI has often been observed. One possible explanation can be that SOI is an Indo-Pacific phenomenon and recorded closer to Mauritius while Niño 3.4 is recorded further away from Mauritius. Furthermore, precipitation may respond better or faster to atmospheric fluctuations than to changes in SST. Thus SOI may be a better index for the purpose of this study. The fact that there is a correlation between winter SOI and precipitation anomaly in the 1961-1990 series but no correlation with the 1991-2012 or 1961-2012 series suggest that the series may be inhomogeneous.

TABLE I. CORRELATION COEFFICIENT BETWEEN ENSO INDICES AND PRECIPITATION ANOMALY AT VACOAS. NUMBERS IN BRACKET ARE THE P-VALUES

Index	Period	Winter	Summer
Niño 3.4	1961-1990	0.237 (0.208)	0.031 (0.875)
	1991-2009	0.065 (0.797)	0.296 (0.219)
	1961-2009	0.161 (0.276)	0.075 (0.610)
SOI	1961-1990	−0.329 (0.076)	−0.015 (0.941)
	1991-2012	0.042 (0.858)	−0.348 (0.113)
	1961-2012	−0.184 (0.198)	−0.118 (0.411)

B. *Correlation between Precipitation and IOD*

Table II gives the correlation between average IOD index and average precipitation anomaly for winter and summer. The correlation for three different periods was computed. A positive, statistically significant, correlation exists between average winter IOD and precipitation anomaly at Vacoas.

TABLE II. CORRELATION COEFFICIENT BETWEEN IOD INDEX AND PRECIPITATION ANOMALY AT VACOAS. NUMBERS IN BRACKET ARE THE P-VALUES

Index	Period	Winter	Summer
IOD	1961-1990	0.375 (0.041)	0.108 (0.572)
	1991-2009	0.210 (0.388)	0.135 (0.482)
	1961-2009	0.309 (0.031)	0.124 (0.394)

IV. PRECIPITATION PREDICTION USING NEURAL NETWORKS

C. *Methods*

An artificial neural network (ANN) is an approach inspired by the biological neural networks in the brain. Through a connectionist approach, it aims to process information, through a collection of nodes, called artificial neurons. ANNs are employed in modeling the relationships between sets of inputs and outputs, often related through complex non-linear systems. In essence, they are non-linear statistical modeling tools.

An ANN is an adaptive system that changes its structure (like weights between connections) based on information provided to the network during a learning phase. It can then be used for the prediction of outputs given new input data.

Artificial neural networks normally consist of three layers: the input layer, the hidden layer, and the output layer. The input layer simply accepts inputs from the external environment, while the output layer presents the results obtained to the external environment. Hidden layers are various levels of transformation of the data to establish the connection between inputs and outputs. The number of hidden layers varies; hidden layers may even be absent in certain types of neural networks.

An ANN has i number of inputs (x_i) and an output y. A weight (w_i) is associated with each input. Often, an additional parameter (w_0), called a bias, is included. The single neuron is a feed-forward device, that transforms a given input into some output. The activation, a, of the neuron is calculated according to the inputs x:

$$a = \sum_i w_i \, x_i \qquad (2)$$

Then, the output y is set as a function $f(a)$. Back-propagation is the most widely used training algorithm [6]. It is a form of supervised learning, where the neural network is given a set of training examples. Each example consists of a pair of input and output and the input is typically represented by a series of values. The training set is used to iteratively update the weights of the interconnections so as to minimize an error function.

ANNs have been successfully used in a variety of applications such as medical diagnosis, vehicle control, recognition of faces and objects, e-mail spam filtering, game-playing, handwritten recognition, automated trading systems, etc. They have also been applied to a wide range of problems in hydrology, for example rainfall-runoff and streamflow forecasting [7, 8], flood and drought prediction [9, 10], etc.

An exhaustive review of the application of ANN in hydrology and water resources engineering is available in [11]

D. *Predicting Precipitation*

We have used an approach proposed in [12] that successfully used ANN for long range precipitation forecasting in California, USA. We started our tests trying to predict the yearly average precipitation, and then seasonal and monthly precipitation. We devised a series of tests to predict the seasonal average precipitation.

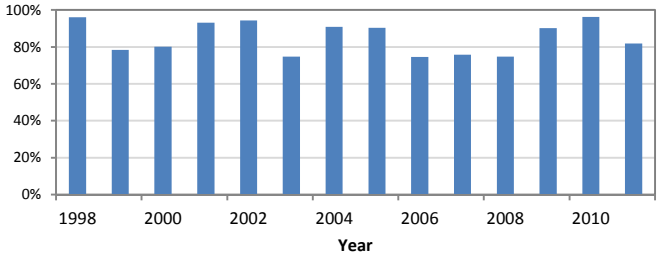


Fig. 5. Percentage accuracy of average winter precipitation using SOI.

For all the tests, 33 to 35 years of data have been used for training, 7 years of data has been used for validation, and 7 years of data have been used to test the neural network. Fig. 4 shows a representative result for the tests cases: the accuracy with which each year's average winter precipitation was predicted using SOI. The accuracies range from 74.8% to 96.2%, with an average of 86%. In general, we found that we could predict winter precipitation fairly accurately, while the prediction for summer are not as accurate. This corroborates with our statistical analysis.

V. DISCUSSION

The ability to accurately predict precipitation and drought conditions is a critical factor in the short, medium, and long-term water resources management. Precipitation is known to be influenced by large-scale circulation patterns. In

Mauritius, two major teleconnection patterns that are likely to affect rainfall are ENSO and IOD.

In this study, our focus was to assess the influence of these circulation patterns on precipitation on rainfall data recorded at Vacoas. The indices used are the Niño 3.4 index, the Southern Oscillation Index, and the Indian Ocean Dipole Mode Index for the period 1961 to 2012.

Our initial analysis revealed poor correlation between precipitation and these climate indices. However, when the data series were split into two separate seasons (winter and summer), higher correlations were obtained for winter. We then predicted precipitation using teleconnection indices as predictor variables. With neural networks, we obtained an average winter precipitation prediction accuracy of 86%. However, prediction of summer precipitation was less accurate.

Our study was based on only one meteorological variable at one station, which is insufficient for accurate modeling and forecasting. Potential future work includes extending the analysis spatially to cover the entire island, using longer time series, and including other variables (such as temperature, pressure, evaporation, and humidity).

REFERENCES

[1] United Nations "International Decade for Action Water for Life 2005-2015: Water scarcity", 2011.

[2] D. Deepchand and C. P. Khedun, "Water Resources Development and Management - Present Status and Future Program", in Development of a Mauritius Country Framework for Action within the Southern African Vision for Water, Life, and Environment in the 21st Century, edited by E. Seenyen, C. Seenyen and C. P. Khedun, Pailles, Mauritius, 2005.

[3] C. P. Khedun, K. Gopal, and A. Sohun, "Modeling the Influence of Large-Scale Circulation Patterns on Precipitation and a Multivariate Drought Analysis for Mauritius", Technical Report submitted to the Mauritius Research Council, 2013.

[4] A. G. Barnston, M. Chelliah, and S. B. Goldenberg, "Documentation of a highly ENSO-related SST region in the equatorial Pacific: Research Note", Atmosphere-Ocean, 35(3), 367 – 383, 1997.

[5] K. E. Trenberth, "The definition of El Niño" Bulletin of the American Meteorological Society, 78(12), 2771-2777, 1997.

[6] H. Maier and G. Dandy, "Neural networks for the prediction and forecasting of water resources variables: a review of modeling issues and applications", Environmental Modelling & Software, 15, 101-124, 2000.

[7] F. Anctil, C. Perrin, and V. Andréassian, "Impact of the length of observed records on the performance of ANN and of conceptual parsimonious rainfall-runoff forecasting models", Environmental Modelling & Software, 19, 357-368, 2004.

[8] C. M. Zealand, D. H. Burn, and S. P.Simonovic, "Short term streamflow forecasting using artificial neural networks", Journal of Hydrology, 214, 32-48, 1999.

[9] A. Mukerji, C. Chatterjee, and N. S. Raghuwanshi, "Flood forecasting using ANN, neuro-fuzy, and neuro-GA models", Journal of Hydrologic Engineering, 14, 647-653, 2009.

[10] A. K. Mishra, V. R. Desai, and, V. P. Singh, "Drought forecasting using a stochastic and neural network model", Journal of Hydrologic Engineering, 12, 626-638, 2007

[11] R. S. Govindaraju and R. S. Rao, "Artificial Neural Networks in Hydrology", Springer Netherlands, 2000.

[12] D. Silverman and J. A. Dracup, "Artificial neural networks and long-range precipitation prediction in California", Journal of Applied Meteorology, 39(1), 57-66, 2000.

Forensic Profiling of Cyber-Security Adversaries based on Incident Similarity Measures Interaction Index

Victor R. Kebande
DigiFoRs Research Group
Dept of Computer Science.
University of Pretoria
Pretoria, South Africa
vickkebande@gmail.com

Nickson M. Karie
Dept. of Computer Science
University of Swaziland
Kwaluseni, Kingdom of eSwatini
nickson.karie@gmail.com

Ruth D. Wario
Department of Computer Science & Informatics
University of the Free State
South Africa
warioRD@ufs.ac.za

H. S.Venter
DigiFoRs Research Group
Dept of Computer Science.
University of Pretoria
Pretoria, South Africa
heinventer@gmail.com

***Abstract*—In today's complex cyber space, forecasting the likelihood or probability that a Cyber Security Adversary (CSA) is likely to attack a given infrastructure, system or a networked environment requires a critical analysis of digital data that at that particular time is treated as potential evidence. Digital forensic tools have more often than not been employed in such tasks, however, this aspect has often faced a number of uncertainties. This paper addresses the lack of effective techniques of profiling CSAs in order to discover adversarial motives based on incident similarity measure metrics. The authors of this paper propose an approach that uses an Incident Similarity Measure Interaction Index (ISMII) metric, through which, for example, two independent Cyber Security Incidents (CSI) can be measured and be correlated in order to link a digital crime to the perpetrator. By realising such measures using the ISMII metric, digital forensic investigators are able to profile, predict, and correlate CSI patterns with a degree of certainty. The result of the study depicts a new ISMII metric that is able to compute closely matching cyber-security based incidents.**

Keywords—forensic profiling, cyber-security, adversary, similarity measure, ISMII

I. INTRODUCTION

Just like any other traditional crime, cybercrime isn't a victimless crime. Profiling an individual, a process or even a bot behind a cybercrime can be a useful component in the forensic investigation process. Profiling as stated by [1], involves the analysis of personal characteristics or behavioral patterns, which allows an investigator to make generalizations about a person or a crime scene. Additionally, profiling also employs analysis to determine whether a particular person may be engaged in a particular crime, as determined by existing evidence. This process thus, requires a critical analysis of existing digital traces that at any particular time may also be treated as potential digital evidence.

Nevertheless, a research paper in [2] state that, at a practical security and forensic level; profiling can help profile cyber criminals during and after an attack, and at a theoretical level, it can allow an individual to build improved threat models [2].

Similarly, research by [3], argue that, the idea that an individual committing crime in cyberspace can fit a certain profile may seem far-fetched, but evidence suggests that certain distinguishing characteristics do regularly exist in cyber criminals. It is for this reason that the authors in this paper were motivated to undertake a research in forensic profiling of cyber-security adversaries based on Incident Similarity Measures Interaction Index (ISMII).

This paper thus proposes an approach that uses an ISMII metric, through which, two independent security incidents can be measured and correlated in order to link a digital crime to the perpetrator. By realising such measures using the ISMII metric, digital forensic investigators are able to profile, predict, and correlate security incident patterns with a degree of certainty.

The paper is structured as follows: Section II covers the background while section III handles related work on ISMII. After this, Section IV explains the proposed ISMII metric. This is followed by Section V that presents evaluations. Finally, the paper concludes in Section VI and make mention of the future work.

II. BACKGROUND

This section introduces the background study on the following parameters: Digital Forensics (DF), Cyber Security Adversaries (CSAs), security incidents and Similarity Measures (SM). DF is discussed to show the essence of post-event response during Digital Forensic Investigations (DFI) while CSAs explore the intent of most attackers. Security incidents bring the reader up to date on matters of security incidents and finally the SM used in this paper is discussed to show what can be extracted from possibly identical incidents.

A. Digital Forensics

Digital Forensics (DF) is considered the branch of forensic science dealing with the recovery and investigation of material found in digital devices, often in relation to digital crimes [21],[22]. In any digital forensic investigation process, the potential digital evidence captured during investigation is arguably one among the most important artefacts of an investigation process [4]. Nevertheless, the digital forensic investigation processes have been described in the ISO/IEC 27043:2015 standard and covers information technology, security techniques, incident investigation principles and processes [5].

Since its inception as mentioned by [6] DF has continued to penetrate the society. Today, DF has become a relevant technology that can be used to obtain digital evidence that can be used in a court of law during criminal or civil proceedings [7]. This involves using scientifically proven methods to collect, acquire, examine, identify, analyse and present digital evidence from digital sources that may have reliable information that may be used to prove or disprove a hypothesis about a security incident [8].

Research has also shown that, as technological trends in DF keeps changing, new challenges are also constantly introduced into the domain. For this reason, approaches need to be developed in DF with the ability to effectively assist during any investigation process. Furthermore, the requirement for such approaches in DF is exceptionally important both for the advancement of the field and for the effective digital investigation process as mentioned by [9].

B. Cyber-Security Adversaries

In modern complex computer networks, security professionals are always faced by the fact that they must defend their organizations against various kinds of security threats. This is mostly driven by the fact that, understanding the different motivations of different Cyber Security Adversaries (CSAs), can help security professionals with threat hunting.

This is echoed by [10] who argues that, when threat hunting, one needs to understand the type CSAs you are facing. While their techniques may be very similar, what motivates them can vary. Understanding what motivates the CSAs can lead to a better understanding of where and when a cyber attacker may strike or when an unwitting accomplice takes measures that present undue risk to the organization [10].

As cyber threats and attacks evolves, it is becoming hard for security professionals to find the appropriate measures to deal with CSAs, the onus is in all of us to be better prepared so we can prevent attacks and respond quickly and appropriately when there is a breach. It is for this reason that this paper proposes an approach that uses an ISMII metric, through which, two potential independent security incidents can be measured and be correlated in order to link a digital crime to the perpetrator. By discovering such measures using the ISMII metric, digital forensic investigators are able to profile, predict, and correlate security incident patterns with a degree of certainty.

C. Cyber-Security Incidents

A Cyber-Security Incident (CSI) may simply mean a warning to an organization that there may be a threat to their information or computer security. Such a warning could also be that a security threat has already occurred in the organizations network.

However, according to [18], a CSI is the violation of an explicit or implied security policy. This may include but not limited to:

- Attempts from unauthorized sources to access systems or data.
- Unplanned disruption to a service or denial of a service.
- Unauthorized processing or storage of data.
- Unauthorized changes to system hardware, firmware, or software.

For this reason, organizations need to have a well-defined plan for handling CSIs as well as for the effective operation of their information systems. In this paper, thus, a new approach that uses an ISMII metric, through which, two independent possible CSIs can be measured and correlated in order to link a digital crime to the perpetrator is proposed.

D. Similarity Measures

Similarity Measures (SM) are metrics which are used in different domains to show the measure of how much alike two data objects are. SM can also essential in solving various pattern recognition problems such as classification and clustering as mentioned by [19]. Research on SM by [20] has employed SM on vulnerability assessment using VSS metrics. In this study, the authors have employed have employed the Euclidean Distance Metric (EDM) and Manhattan Distance Metric (MDM) to test the SM_1 and SM_2 as is shown in equation 1 and 2 respectively.

$$SM_1 = \sqrt{\sum_{i=1}^{p} (V_{ij} - V_{jk})^2} \quad (1)$$

$$SM_2 = \sum_{i=1}^{n} | V_{ij} - V_{jk} | \quad (2)$$

In the context of this paper, several SMs have been proposed in literature to compute the SM indexes between different data objects, however, none of the existing SMs have attempted to propose an ISMII metric in the way that is discussed in this paper. This makes existing SM to have a deficiency in profiling CSAs as presented in this paper.

Note that, when it comes to profiling CSAs, measuring the SM between different criminal activities can be an important operation in aiding digital forensic investigations. Hence, this paper proposes an ISMII for profiling CSAs. The ISMII concept is a basic concept that also considers human cognition. The next section explores work that has been used as related work.

III. RELATED WORK

This section gives a highlight of works that have been used as related work. Research by [9], has explored how the semantic similarity between DF terminologies can be measured using web search engines. The work by these authors gave a proposition of Digital Forensic Absolute Semantic Similarity (DFASSV) which employed one-dimensional Euclidean distance for the SM of DFASSV. This work mainly focused on the ever changing DF terminologies. Also, research work by [11],[12],[13] employed Minkowski Distance Metric (MDM) function to check forensic event SM of Potential Security Events (PSEs). This work mainly focused on finding the root square differences for PSEs between set of attributes, finding the absolute difference between the pair of event attributes and examining the absolute difference in magnitude between set of event attributes. These have been given by equation 3, 4 and 5 respectively.

$$d(w_{ij}) = \sum_{k=1}^{n} |W_{ik} - W_{jk}| \qquad (3)$$

$$d(W_{1TP}, W_{2TP}) = \sqrt{\sum_{i=1}^{n} (W_{1TP} - W_{2TP})^2} \qquad (4)$$

$$d(W_{1TP}, W_{2TP}) = Max_i |W_{1TP} - W_{2TP}| \qquad (5)$$

Also, research by [14] has proposed a forensic criterion for matching an assessment of two dimensional multimedia content for judicial disputes concerning graphic symbol similarity. This research applied Histogram Euclidean Distance for two distinct picture histogram, Histogram intersection distance and the histogram quadric distance to give the correlation between histogram bins based on perceptual similarity. This has been achieved using equation 6, 7 and 8 respectively.

$$d_{HE}(H_1, H_2) = \sum \sqrt{(H_1(x,y,z) - H_2(x,y,z)^2} \qquad (6)$$

$$d_{H1}(H_1, H_2) = \frac{\sum_{x\in X, y\in Y, z\in Z} \sqrt{(H_1(x,y,z), H_2(x,y,z)^2}}{[\sum_{x\in X, y\in Y, z\in Z} H_1(x,y,z), H_2(x,y,z)]^2} \qquad (7)$$

$$d_{HQ}(H_1, H_2) = \sum \sqrt{(H_1 - H_2)} (H_2 - H_2) \qquad (8)$$

The next section explains the proposed ISMII metric in this paper.

IV. INCIDENT SIMILARITY MEASURE INTERACTION INDEX (ISMII)

This section gives a description on the approach that the authors have used to profile CSAs based on the Incident Similarity Measure Index (ISMII) as a contribution. Formally, it is worth noting that the concept of ISMII attempts to show the degree of likeness between two distinct CSIs for purposes of profiling CSAs. The authors have given definitions and properties that have been employed in this section as follows:

Definition 1: *A network environment is represented as a graph* $G = (V, E)$ *where* V_i *is a vertex and E is a set of edges* $\{e_1, e_2, e_3, \ldots\ldots\ldots e_n\}$, *which can also be represented as a network node.*

Definition 2: *Interaction index: Formally, this can be represented as a cooperation between two entities that formally can be represented as* $s(i,j) = v\{i,j\} - v(\{i\}) - v(\{j\})$ *where there exists a synergy for* i *and* j *respectively* [15].

Definition 3: *Incident Similarity Measure Index (ISMI) for a network between the nodes can be represented as a Tuple*(G, Pk), *where* $G = (V, E)$ *and pk is the interaction between two CSIs or nodes n, where* $n \in V$.

In order to provide an effective cyber-adversary profiling approach, the authors propose a technique of excavating complex adversarial patterns based on $G = (V, E)$using the ISMII. This helps to show the likeness or relatedness of, for example, two entities or security incidents. By relying on the ISMII, it is possible to discover new adversarial targets, which can also help to predict futuristic attacks.

Given two distinct CSIs I_{1S} and I_{2S}, from a network node that is represented as $G = (V, E)$ can be defined based on the ISMII and the set of CSIs I_{1S} and I_{2S} respectively, which is represented as a weighted Graph (WG). This implies that given I_{1S} and I_{2S}, $WG = (I_{1S} U I_{s2})$. On the same note, the degree of resemblance can be computed based on the average distance and Euclidean distance as shown in Equation 1 and 2 respectively. The authors have been motivated to use the two Similarity Measure (SM) metrics because Euclidean is a common SM metric and very easy to compute [15] while average distance SM metric can easily be used independently while measuring the CSI resemblance. Equation (9) and (10) shows one-dimensional SM equations that can be used to match two specific entities given that there exists a distance between them.

$$d_{av} = (\frac{1}{n} \sum_{i=1}^{n} (I_{1S} - I_{s2})^2)^{1/2} \qquad (9)$$

$$\boldsymbol{d_{Euc} = [\sum_{i=1}^{n} (I_{1S} - I_{s2})^2]^{1/2}} \qquad (10)$$

In the next section an overview of how the ISMII is modeled is given and an explanation follows further on.

A. Modeling ISMII (Overview)

This section presents approaches for modeling ISMII for purposes of realizing effective cyber-adversary profiling. The authors envisages that ISMII is realised given that there exist a networked environment where information is transmitted from source node to destination.

B. ISMII Model

The key elements for ISMII include NE, which has been represented as a networked environment, that is represented as $G = (V, E)$. Through the NE, the nodes $\{e_1, e_2, e_3, \dots\dots\dots e_n\}$ are able to be accessed S.T$\{e_1, e_2, e_3, \dots\dots\dots e_n\} \in NE$. Nevertheless, ISMII also comprise of an object, NE_{obj} S.T$NE_{obj} \in NE$. This object is represented as a potential CSI source S.T $NE_{obj} \in NE \rightarrow ACTN$; where $ACTN$ is an action that can yield Potential Security Incidents, I_s. On the same note, ISMII consist of a cyber-forensic database, $Dbase$, which stores potential and suspicious CSIs or events like IP address (IP), malicious links (L), suspicious files, Metadata (M) and generated Hash values (Hsh). These information can be represented as follows:

NE→ACTN{Dbase{⟦IP⟧_(n+1)∪M_(n+1)∪L_(n+1)∪⟦Hsh⟧_(n+1)} where

$$IP_{n+1} = \{IP_1, IP_2, \dots\dots\dots IP_{n+1}\} \quad (11)$$
$$M_{n+1} = \{M_1, M_2, \dots\dots\dots M_{n+1}\} \quad (12)$$
$$L_{n+1} = \{L_1, L_2, \dots\dots\dots L_{n+1}\} \quad (13)$$
$$Hsh_{n+1} = \{Hsh_1, Hsh_2, \dots\dots\dots Hsh_{n+1}\} \quad (14)$$

As a result,

- $+ IP_{n+1}$ ($IP_1, IP_2, \dots\dots\dots IP_{n+1}$) means that if an adversary's IP address is located, it is flagged if the degree of sameness holds, this shows that $IP_{n+1} \in NE$.
- $+ M_{n+1}$ ($M_1, M_2, \dots\dots\dots M_{n+1}$) means that if an adversary's IP address is located, it is flagged if the degree of resemblance holds, this shows that $M_{n+1} \in NE$.
- $+ L_{n+1}(l_1, L_2, \dots\dots\dots L_{n+1})$ means that if an adversary's IP address is located, it is flagged if the degree of likeness holds, this shows that $L_{n+1} \in NE$.
- $+ Hsh_{n+1}(Hsh_1, Hsh_2, \dots\dots\dots Hsh_{n+1})$ means that if an adversary's IP address is located, it is flagged if the degree of sameness holds, this shows that $Hsh_{n+1} \in NE$.

Based on the similarity interaction index that is realized as a result of the above (Equation 11-14) then

$$IP_{n+1} \cup M_{n+1} \cup L_{n+1} \cup Hsh_{n+1} \in NE = ISMII \quad (15)$$

Notably, two distinct adversarial CSIs I_{s1} and I_{s2} that holds closely related ISMII can be profiled If a third adversarial CSI I_{s3} exhibits closely matching attributes to I_{s1} and I_{s2} respectively, i.e $I_{s1} \cup I_{s2} \rightarrow I_{s3}$ which shows that there exist a relationship between CSIs S.T

$I_{s1} \rightarrow\cup I_{s2}, I_{s2} \rightarrow\cup I_{s3}$, therefore $I_{s1} \rightarrow\cup I_{s2} \rightarrow I_{s3}$ (16)

Given that these CSIs are exhibited in an NE that is represented by $G = (V, E)$, then Figure 1 shows the relationship between nodes 1-5 and the potential sources of adversarial CSIs. Based on ISMII, then NE can be represented as follows:

$$NE(1,2,3,4,5) = \{ I_{s1} \rightarrow\cup I_{s2} \rightarrow I_{s3}\} \quad (17)$$

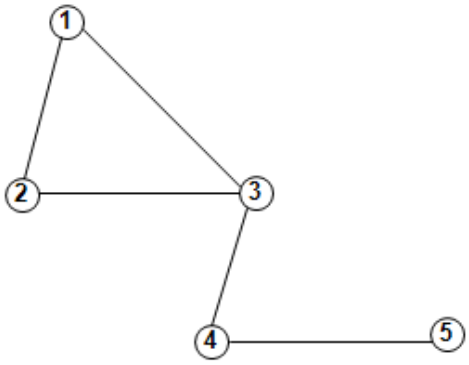


Figure 1. Collaboration between nodes in a networked environment

Therefore, using ISMII approaches, the following relationship can be used to infer the potential sources of profiling CSAs. Therefore, using ISMII approaches, the following relationship can be used to infer the potential sources of profiling based on $NE(1,2,3,4,5) = \{ I_{s1} \rightarrow\cup I_{s2} \rightarrow I_{s3}\}$. For each, the connectivity of nodes of Figure 1 is shown based of the potential attributes of CSIs $\{I_{s1} \rightarrow\cup I_{s2} \rightarrow I_{s3}\}$.

$$NE(1,2) \rightarrow I_{s1}\{IP_{n+1} \cup M_{n+1} \cup L_{n+1} \cup Hsh_{n+1}\}, I_{s1} \in NE$$
$$NE(1,3) \rightarrow I_{s1}\{IP_{n+1} \cup M_{n+1} \cup L_{n+1} \cup Hsh_{n+1}\}, I_{s1} \in NE$$
$$NE(1,4) \rightarrow I_{s1}\{IP_{n+1} \cup M_{n+1} \cup L_{n+1} \cup Hsh_{n+1}\}, I_{s1} \in NE$$
$$NE(1,5) \rightarrow I_{s1}\{IP_{n+1} \cup M_{n+1} \cup L_{n+1} \cup Hsh_{n+1}\}, I_{s1} \in NE$$
$$NE(1,2) \rightarrow I_{s2}\{IP_{n+1} \cup M_{n+1} \cup L_{n+1} \cup Hsh_{n+1}\}, I_{s2} \in NE$$
$$NE(1,3) \rightarrow I_{s2}\{IP_{n+1} \cup M_{n+1} \cup L_{n+1} \cup Hsh_{n+1}\}, I_{s2} \in NE$$
$$NE(1,4) \rightarrow I_{s2}\{IP_{n+1} \cup M_{n+1} \cup L_{n+1} \cup Hsh_{n+1}\}, I_{s2} \in NE$$
$$NE(1,5) \rightarrow I_{s2}\{IP_{n+1} \cup M_{n+1} \cup L_{n+1} \cup Hsh_{n+1}\}, I_{s2} \in NE$$
$$NE(1,2) \rightarrow I_{s3}\{IP_{n+1} \cup M_{n+1} \cup L_{n+1} \cup Hsh_{n+1}\}, I_{s3} \in NE$$
$$NE(1,3) \rightarrow I_{s3}\{IP_{n+1} \cup M_{n+1} \cup L_{n+1} \cup Hsh_{n+1}\}, I_{s3} \in NE$$
$$NE(1,4) \rightarrow I_{s3}\{IP_{n+1} \cup M_{n+1} \cup L_{n+1} \cup Hsh_{n+1}\}, I_{s3} \in NE$$
$$NE(1,5) \rightarrow I_{s3}\{IP_{n+1} \cup M_{n+1} \cup L_{n+1} \cup Hsh_{n+1}\}, I_{s3} \in NE$$

Based on the equations above on profiling sources $(1,2,3,4,5) = \{ I_{s1} \rightarrow\cup I_{s2} \rightarrow I_{s3}\}$, the probability that a sequence of CSIs in a NE will provide similar or closely related adversarial CSIs will depend on $I_{s1} \rightarrow\cup I_{s2} \rightarrow I_{s3}\} \in NE$. If there exist ISMII, then the CSIs can be profiled and then sent to the database as shown in Figure 2.

Based on the equation above on profiling sources, the probability that a sequence of CSIs in a NE will provide similar or closely related adversarial CSIs will depend on $I_{s1} \rightarrow\cup I_{s2} \rightarrow I_{s3}\} \in NE$. If there exist ISMII, then the CSIs can be profiled and then sent to the database as shown in Figure 2.

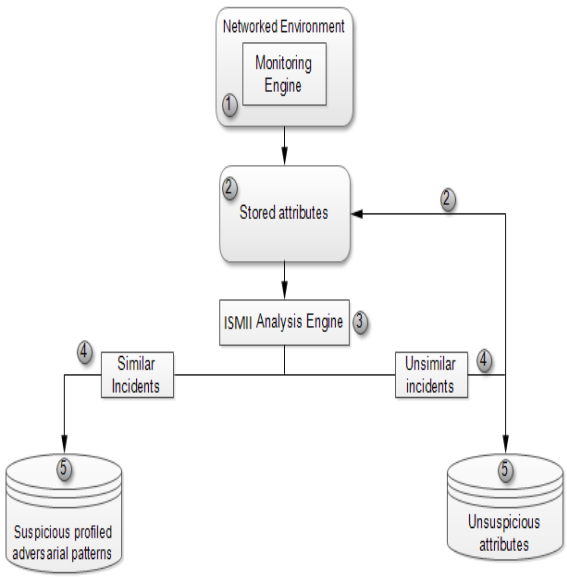


Figure 2. Adversarial profiling model

Based on Figure 2, a monitoring engine in the step labeled 1 in the networked environment is able to monitor the file attributes being transmitted. These files can either *be IP address, metadata* or *hash values*. In the section labeled 2, the previously stored unsuspicious files are accessed and then they are subjected to an ISMII, which is performed by ISMII Analysis Engine (ISMII-AE) in step 3. ISMII-AE determines if there exist some similarity among the attributes being transmitted. If some similarity is identified in step 4, then the adversarial CSI is recorded and flagged as suspicious and this CSI is then stored in suspicious attribute database, otherwise it is stored in unsuspicious attribute database as is shown in Step 5.

In our research, we assume that two distinct CSIs that are represented by $I_{is} = \{I_{is1}, I_{is2}, \ldots\ldots I_{isn+1}\}$ and $I_{ip} = \{I_{ip1}, I_{ip2}, \ldots\ldots I_{ipn+1}\}$ where I_{is} and I_{ip} represents CSIs that exhibits some distance that is likely to show some degree of closeness or similarity. In order to determine the ISMII between I_{is} and I_{ip} we compare the set of CSIs based on the following SM metric functions in equation (9) and (11) respectively, the generated ISMII for the CSI is shown in equation (18).

$$ISMII = d(I_{s1}, I_{s2}) = \sum_{i,j} I_{s1}, I_{s2} \quad (18)$$

where $i,j>0$ shows the importance of matched CSI, where there can exist n number of SCIs. Using this formulae we are able to define a technique through which a CSI that matches a given descriptor or the pattern of occurrence corresponds to a known attack.

Given that our SM is used to pertinently check the closeness or how the CSI match, it is important to note that the choice of our SM is robust for even randomly selected CSIs since the comparison lies between I_{is} and I_{ip}, two distinct CSIs.

V. EVALUATION

The set of study that has been conducted in this research paper has highlighted approaches that can be used while conducting CSAs profiling. The mathematical representation that has been used as a basis for ISMII model, together with other propositions presents the contribution of this paper. The approaches have been carefully presented in order to highlight significant forensic approaches that are able to filter suspicious patterns towards CSI identification. To the best of the authors' knowledge, this proposition provides a technique that also relies on incidental characteristics on a given network topology given the representation of $G = (V, E)$, in order to realise a proactive-based approach. By relying on the degree of incidental-sameness security of networks can be guaranteed to a degree of certainty too. It is also worth to note that, the approach that has been proposed in this context can easily be extended to different distinct attack scenarios. This can also enable digital forensic investigators to apply the IISMII metric to determine or excavate evidence based on adversarial patterns experiences at that the NE nodes.

VI. CONCLUSION

In this paper, the authors have introduced ISMII approach that has been used to profile cyber-security adversaries, which in the long run realizes a proactive-based technique of hardening network security through early detection. The authors have also modeled ISMII technique with a focus of bring a suitable similarity measure on CSIs. Our future work will be based on practical applicability of ISMII through a prototype through better shaping of incidental likeness in a cyber-security perspective.

ACKNOWLEDGMENT

The authors gratefully acknowledges the support of Department of Computer Science, and the Information and Computer Security Architecture (ICSA) and DigiFoRs Research group, University of Pretoria, Department of Computing and Informatics, University of The Free State for support while coming up with this research. It is worth noting that any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the University of Pretoria or the University of The Free State.

REFERENCES

[1] R. Yepes. The Art of Profiling in a Digital World," The Police Chief, 2016. Available at: http://www.policechiefmagazine.org/the-art-of-profiling-in-a-digital-world/?ref=45859a8c28c48f4a72715597941e2220 [Accessec, 07/07/2018]

[2] A. Filippoupolitis, G. Loukas, S. Kapetanakis. Towards real-time profiling of human attackers and bot detection. Canterbury Christ Church University, UK, 2014.

[3] N. Nykodym, R. Taylor, J. Vilela. Criminal profiling and insider cyber crime. Digital Investigation. Vol 2, Issue 4, Dec. 2005, Pages 261-267, 2005.

[4] Y. Yusoff, R. Ismail, Z.Hassan. Common Phases of Computer Forensics Investigation Models. International Journal of Computer Science & Information Technology (IJCSIT), Vol. 3, No. 3, 2011,

[5] ISO/IEC 27043:2015 Preview Information technology -- Security techniques -- Incident investigation principles and processes. Accessed at https://www.iso.org/standard/44407.html

[6] N.M.Karie, V.R. Kebande H.S.Venter. A Generic Framework for Potential Digital Evidence Traceability. In the Proceedings of the EWWCS – 2016, Germany.

[7] V.R. Kebande, N.M Karie. A Generic Digital Forensic Readiness Model for BYOD using Honeypot Technology-IST Africa , 2016.

[8] B.D.Carrier, E.H.Spafford. Defining Digital Crime Scene Event Reconstruction, The Journal of Forensic Sciences 49(6), 2004.

[9] N.M.Karie, H.S Venter. Measuring Semantic Similarity Between Digital Forensics Terminologies Using Web Search Engines. In the Proceedings of the 12th Annual Information Security for South Africa Conference. Johannesburg, South Africa. Published online by IEEE Xplore®, (pp. 1-9), 2012.

[10] B. Metivier.Threat Hunting: Six Cyber Adversaries to Pursue. Available at: https://www.sagedatasecurity.com/blog/threat-hunting-six-cyber-adversaries-to-pursue [Accessed, 1st Sept, 2018], 2018.

[11] V.R.Kebande. A Novel cloud Forensic Readiness Service Model. Upspace, accessed at https://repository.up.ac.za/handle/2263/66140, 2018.

[12] V. R. Kebande, H. S. Venter. Adding event reconstruction to a Cloud Forensic Readiness model. In *Information Security for South Africa (ISSA), 2015* (pp. 1-9). IEEE, 2015.

[13] V.R.Kebande, H.S. Venter. Novel digital forensic readiness technique in the cloud environment. *Australian Journal of Forensic Sciences*, *50*(5), 552-591, 2018.

[14] B. Vasic. New content based forensic metrics for judicial disputes concerning the graphic symbols similarity. In *Telecommunication in Modern Satellite, Cable and Broadcasting Services (TELSIKS), 2013 11th International Conference on* (Vol. 1, pp. 341-344). IEEE, 2013.

[15] P. L. Szczepanski, A.S Barcz, T.P. Michalak, T. Rahwan. The Game-Theoretic Interaction Index on Social Networks with Applications to Link Prediction and Community Detection. In *IJCAI* (Vol. 15, pp. 638-644), 2015.

[16] G. Gan, C. Ma, J.Wu. Data Clustering theory, Algorithms, and Applications. ASASIAM Series on Statistics and Applied. Society for Industrial and Applied Mathematics; 2007.

[17] D. Hand, H.Mannila, P. Smyth. Principles of data mining(adaptive computation and machine learning). Drug safety. 2001.

[18] P.Cichonski. Computer Security Incident Handling Guide. Recommendations of the National Institute of Standards and Technology. Available at: https://nvlpubs.nist.gov/nistpubs/specialpublications/nist.sp.800-61r2.pdf [Accessed 2nd September 2018], 2018.

[19] F.Iglesias, W. Kastner. Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns. Energies, 2013.

[20] V.R. Kebande, I.Kigwana, H.S.Venter, N.M. Karie, , D.R.Wario CVSS Metric-Based Analysis, Classification and Assessment of Computer Network Threats and Vulnerabilities 2010 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD), 2018.

[21] M. Reith, C.Carr, G. Gunsch. An examination of digital forensic models. *International Journal of Digital Evidence*, *1*(3), 1-12, 2002.

[22] B. Carrier. Defining digital forensic examination and analysis tools using abstraction layers. *International Journal of digital evidence*, *1*(4), 1-12, 2003.

Delivery of an Interactive Audio Course on Fisheries Law via Dumb Phones: Agalega Island as a Case Study for Testing a Novel Distance Education Platform - The Audio MOOC

Raj Kishen Moloo
Department of Digital Technogies, University of Mauritius(UoM)
Reduit, Mauritius
r.moloo@uom.ac.mu

Kavi Kumar Khedo
Department of Digital Technogies, University of Mauritius(UoM)
Reduit, Mauritius
k.khedo@uom.ac.mu

Tadinada Venkata Prabhakar
Department of Computer Science, Indian Institute of Technology Kanpur (IITk),
Kanpur, India
tvp@iitk.ac.in

***Abstract*—This work is a continuation of a larger research work which advocates that Distance Education (DE) through audio-only learning mode can be developed into a full fledge audio-MOOC. Audio MOOC framework is an innovative framework which enables learning through mere phone calls. It has been conceived to digitally include low literate population in the education process by opening up access to learning materials to the unreached and the have-nots usually hindered by barriers such as language, literacy, culture, connectivity and distance which existing MOOCs have failed to address. This work demonstrates how our proposed framework is used to connect to a remote island lost in the middle of the Indian Ocean with limited maritime and air access but which since some few years back can be connected via basic phones through voice calls. Agalega is an ideal test case scenario for our research since it characterizes remoteness, limited connectivity, semi-literate population with limited access to education which our research aims at addressing. A group of 50 Fishermen was identified from both the Agalega islands. The course was of 9 days duration from 15 to 23 September 2017. The system was conducted live over the telephony network making use of our GSM gateway. The specificity of the system was that our GSM gateway resided in Mauritius connected to a cloud server, while the course was delivered to people of Agalega 1100 Km far from Mauritius over the sea. Nevertheless, our system performed as expected and proved to be a success.**

Keywords—Audio MOOC, Audio Learning System, Distance Education for low-literate population, Mobile for Development (M4D)

I. INTRODUCTION

Agalega island is an atoll situated at about 1,100 Km North of Mauritius and 563 Km South of Seychelles. It forms part of the Mauritian territory and is made up of two small islands, the North and the South which are separated by a narrow stretch of about 2 km channel formed by reefs on two sides as shown in figure 1. The North Island is 12.5 km long and 1.5 km wide, while the South Island is 7 km long and 4.5 km wide, covering a total area of 26 square kilometers. It has an estimated total population of 300 people. There are two villages in the North Island namely village Vingt Cinq (25) and la Fourche while Ste Rita is the only village on South Island. Vingt Cinq (25) is the centre of administration with La Grande Case being its main administrative headquarter, for both islands [1].

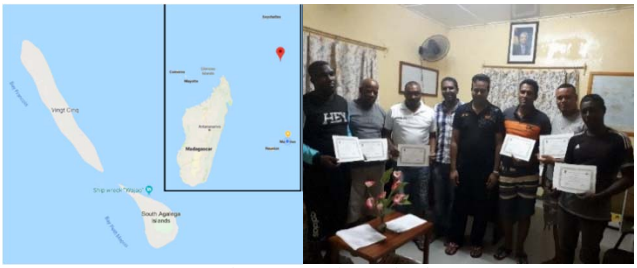


Figure 1. Agalega Island

The Outer Island Development Corporation (OIDC) is a government body which is responsible for the management and development of the island. The main economic activity of the island is coconut oil production and fishing. Access to the island is very limited and is deserved mainly by the Mauritius Trochetia passenger-cargo ship which provision the island with passenger transport, food supplies and necessities every three months. In case of urgency, the Mauritius National Coast Guard sends it patrol aircraft, the Dornier as an air link or else the CGS Baracuda, a waterjet fast patrol vessel [2]. In terms of telecommunication, Agalega North Island was connected to the Mauritius mobile network since 2008 [3] by Emtel Ltd. Since, people from Agalega can make phone calls mainly on the North Island, communication being rather sporadic and localized in Agalega South Island. Only voice calls and SMS are allowed and they are at the same rates practiced in Mauritius. Unfortunately, till date, there is no internet connectivity on the island.

Schooling goes up to lower secondary in Agalega. For higher education, students come to Mauritius where they live with foster parents, relatives as well as in foyers. Among the adult population, schooling is very basic and this explains the low literacy among them. Inhabitants are mostly fishermen and farmers involved in coconut production and processing. Traditions and subsistence skills are transmitted orally. Formal training and empowerment are very sparse and Agalegaens need to come to Mauritius for these programmes. Distance Education programme is quasi-

inexistent, except TV and Radio programmes. But these programmes are general programmes and do not respond to the needs of the island. Being a remote and disconnected island, it is difficult and costly to respond to individual training and capacity building needs for each inhabitant. Literacy among the adult population is very low and usually oral learning is privileged. An interesting development, however, is the advent of mobile phones on the island. Almost everyone has a mobile phone nowadays. Though semi-literate, the adult population knows how to use a mobile primarily for making calls and to a lesser extent SMS.

This makes an ideal test case scenario for the Audio MOOC framework developed by our team. It is an innovative Distance Education (DE) framework which enables audio learning through mere phone calls. It has been conceived to digitally include low literate population in the education process by opening up access to learning materials to the unreached and the have-nots usually hindered by barriers such as language, literacy, culture, connectivity and distance. Agalega fits perfectly well in our research project since it is remote but can be connected by mobile calls and locals are low literate people with limited access to education/ training materials. The specificity of the system was that our GSM gateway resided in Mauritius connected to a cloud server, while the course was delivered to people of Agalega 1100 Km far from Mauritius over the sea.

With the support of OIDC and the Fisheries Training and Extension Centre (FiTEC) Mauritius, a course on Fisheries Law was mounted and deployed on the platform. 50 Fishermen volunteered to test the system by taking the course, out of which 19 successfully completed. Hence, this research work aims at demonstrating how we developed the audio course, deployed and tested our system on the local fishermen. We also present the findings based on the survey we conducted and analytics registered on the system and how the framework can be used to bridge the digital divide and empower remote and deprived population.

II. LITERATURE REVIEW

In such case scenario, as per literature, Distance Education (DE) is usually privileged, being more cost-effective. According to Peters [4], the cost factor has been the main driver in the adoption of DE. It is a form of industrialised teaching and learning which involves the delivery of teaching materials to a high number of students at the same time and anywhere irrespective of where they live, hence overcoming the distance barriers. DE has evolved over the years with the advancement of technology, starting from correspondence courses via postal service to radio, television, E-learning and Mobile Learning.

Even though there is a multitude of DE medium, as per the specificity of Agalega, audio learning is a more feasible and cost-effective solution. Spoken language interface has been argued to be the most logical, acceptable and natural way to interact with a low-literacy population [5] [6] [7] since it does not require literacy. It is an intuitive medium understandable by everyone, irrespective of their cultural background and literacy level, whether textual or technology [8]. Several audio learning technologies have emerged over the years to respond to the education needs of the low literate population. Radios, audio-cassettes, CD/DVD, portable media (MP3 Players) players and mobile phones are some of the technologies which are very easy to use and widely available even in the poorest region around the world.

According to [9], there are several types of audio-based distance education, classified as follows: Broadcast Radio (BR), Interactive Radio Instruction (IRI), Interactive Audio Instruction (IAI), Audio Conferencing (AC) and Call Centre (CC) model. BR is usually used under the following circumstances: where radio is already a common technology; technology deprived area; there is a scarcity of qualified teachers; a large-scale education outreach is needed. A wide range of subject areas, from farming practices [10], rural development [11], enhancing literacy [12] and [13] programmes are disseminated over the medium. IRI makes BR more interactive, emphasizing on active learning and improve teaching practices. It allows students to respond to radio prompts and interact with materials and peers at the radio's prompting [14]. Instruction is scaffolded across a series of episodes. IAI is a low-cost DE which delivers learning materials radio or audio technology such as CD/DVD. It is mainly used to outreach deprived population who do not have access to the conventional education system, including the rural poor and children with disabilities. It has been found to be effective in unstable and conflict-affected regions by extending the reach of broadcast and interactive radio, both of which are highly vulnerable to broadcast interruptions.

According to Cooksun [15], telephone is the simplest and most accessible technologies used in DE to mentor and reach numerous students simultaneously. Audio-conferencing (AC) has long been used in DE to build up learning communities from students scattered in different geographic regions. It is a synchronous technology which allows group of learners to participate via phone calls simultaneously. Macmullen [16] argues that AC provides a high level of interaction and supports a broad range of learning styles. However, it requires a certain level of technical skill to operate. Also, the learners' independence and flexibility are compromised. Another telephony-based learning system is the call centre model which provides instructional services, provide information to prospective students and to answer course-related queries from off-campus students on a day-to-day basis [17]. It has the advantage of being accessible from anywhere and anytime.

A. Existing telephony based learning system

Several systems based on VoIP, IVR and open source PBX technologies have been implemented over the years to dispense educational message. Mobile Kunji and Mobile Academy are examples of IVR Based systems which allows families in Bihar to access health information via simple phone calls. These systems have been supported by the BBC Media Action and Gates Foundation. The Capacity Plus project is a mobile learning system used in Senegal for counseling, managing and to dispel misconceptions of contraceptions. It combines IVR and SMS text messaging

[18]. Learning through Interactive Voice Educational Systems (LIVES) is a push-based voice learning system developed by the British Columbia University. The system calls the students at a scheduled time for course delivery [19]. Voice-enabled interactive service (VòIS) is an E-learning system which uses IVR and voice recognition technology [20]. Blind students from higher education institutions in the USA provided very positive outcome response in the use of VoIS. Garcia et al. [21] propose a "Voice Interactive Classroom", which is a middleware approach for cross-platform and multi-channel access to internet-based learning that can be reused and integrated within various elearning platforms.

III. THE AUDIO MOOC SYSTEM

The Audio MOOC system has been derived from the framework proposed by Moloo, Khedo & Prabhakar [9]. It provides the following features: 1. A simple easy to user interface based on IVR which allows a learner to interact with audio learning contents over the phone via DTMF keys; 2. Navigation keys allowing learners to control the pace of learning through the next, previous and pause buttons; 3. LMS support features to integrate the learning lifecycle on a single platform such as user management, content creation and dissemination, learner tracking, progress and performance monitoring with the possibility of having online assessment and exam. 4. A flexible architecture which allows for customization to respond to organization and learner needs. 5. Audio pedagogy features which considers the cognitive aspects of learners for an enhanced learning experience.

It has a three-tier architecture which are as follows: the Application tier, the Softswitch to Application tier and the Hardware tier. The Application layer provides features of an Audio Learning Management System like User Registration, Course Administration, Assessment/Exam Management, Call Management and Learner Analytics. The Softswitch to Application layer provides VoIP telephony features to the Application Layer. The Softswitch layer allows for call management and is linked with external gateways to bridge GSM and PSTN network. The system components have been implemented modularly as plugins, hence easily extensible and scalable. From a technology perspective, it provides features such as User Management, LMS support, Analytics and Learner tracking capabilities, IVR interfacing capabilities and VoIP configuration functionalities to mention a few. From a learner perspective, the system provides an easy to use IVR interface allowing participants to control their learning pace by using navigation capabilities. A user can connect to a course anytime by dialing a specified number. Users can navigate through the different lessons in the course using features like play/pause, go to the next or previous lesson. At any time he can disconnect. The system registers where he stopped. Next time the user dials again, he is directed from where he stopped previously. From an administrator/ admin perspective, they can easily create and deploy courses and tests with the capability of monitoring their students using the learner analytics feature.

IV. DELIVERY OF THE COURSE ON FISHERIES LAWS ON AIR

A. Audio Lessons Design and Preparation

A presentation was made to the OIDC on the novel platform and how it can help people from Agalega in their educational needs. The OIDC recommended that courses related to fishing activities would be more suitable and appealing to them. The fishermen community are used to traditional fishing practice on the island and most of them are not aware of the do's and don'ts in fishing. Hence, for a start, a course on Fisheries Law was suggested for awareness among the fishermen community. Subsequently, experts from the Fisheries Training and Extension Centre (FiTEC) of Mauritius were contacted for their support.

Audio content creation was a challenge for this project. Courses had to be designed to maximize the learning experience through the auditory medium only. Hence, audio lessons were prepared according to the audio instructional guidelines proposed by [9]. Courses were designed to captivate the learners' attention for an enhanced learning experience. The following pedagogical features were considered: a) human attention span, b) learner cognitive capacity c) content and d) audio recording and voice quality. The course was split into seven (7) chapters based on Miller's Law [22] stating that people can retain 7+-2 objects in working memory. This catered for the learners' pace of learning through the navigation capabilities provided like next, previous and pause buttons. To promote discrimination of new concepts, end segment questions were included at the end of each chapter [23]. The lessons were designed to be of an average of 120 seconds, duration to fit the human attention span [5]. This is confirmed by [24] stating that students are much more engaged within the first three minutes of video lectures. The course development was a tedious process which took around three months (part-time). Many revisions were made to align it with the proposed instructional guidelines. The content was designed to be relevant and concise to fit within the human attention span. It incorporated clearly defined and realistic objectives [25] and Bloom's Taxonomy [26] for higher level thinking.

Once content was finalized, it was time for recording. This is another important process and we followed the auditory considerations proposed by [9]. No background music was included so as not to distract the audience. A voice artist was engaged for proper voice modulation in a conversational style with emphasis on keywords. To avoid monotony, the voice used was modulated in terms of its pace, pitch and volume. To maximize retention of information presented, we used proper pacing and appropriate pauses in-between key pieces of information [27]. Details of the lessons in Fisheries Law course and their respective audio duration are as per Table 1. Seven (7) Audio Lessons excluding an introduction and a conclusion have been created by the FiTEC experts. The total duration of the course was approximately 17 minutes with an average of 112 seconds per audio lessons.

TABLE 1 – AUDIO LESSONS AND THEIR DURATION

Fisheries law in Mauritius	Duration
Introduction	1:34
Law interpretation and definitions	2:57
Prohibited fishing methods and gears	2:22
Fishing methods and corresponding laws	3:11
Protection of Fish	2:25
Undersized fish	1:42
Prohibition of Removal of Corals and Shellfish	1:59
Conclusion	0:41
TOTAL	**16:51**

B. IVR and Navigation Design Consideration

The Interactive Voice Response (IVR) is the only interface provided by the voice medium to interact with audio contents. We used a user-centered approach for its design to make it as simple and intuitive as possible while maintaining its robustness. The best design practice from [9] has been adopted. The Dual Tone Multiple Frequency (DTMF) tones input was preferred compared to speech recognition since it has been found to be more accurate and instantaneous. The IVR design had the following attributes:

1. Options menus have been limited to 5-9 choices to maximize retention according to Miller's Law [22] on the working memory span.
2. Menu Options have been kept within 30 seconds for a more efficient IVR experience.
3. Appropriate pauses between menu items have been included to allow users enough time to press the correct key.
4. Navigation options have been provided. E.g Option to repeat, go to previous, go to next lesson.
5. Options keys have been used consistently.
6. The local language with a friendly tone has been used.
7. No technical terms and jargons were used for a clearer understanding.
8. Short and concise sentences with appropriate speech rate were used for menus and prompts and were done in a turn-taking manner to make it less formal and more conversational.
9. No background music was used.

C. Delivering and Managing the Course

With the help of OIDC, we identified around 50 fishermen, 30 from Agalega North Island and 20 from Agalega South Island willing to follow the course. OIDC provided their names and mobile numbers on an excel sheet which we used in our batch registration module to register them on the system. Prior to the start of the course, we called the fishermen and gave a brief overview of the modalities of the course and how to use the system. Unfortunately, fishermen residing in Agalega South Island were unreachable by phone for known reasons of connectivity issues on the island. Hence only the 30 fishermen from Agalega North participated in the course on Fisheries Law. Age range varied from 20 to 60 years. Each fisherman received a Rs 100 calling card prior to the start of the course. Upon completion, they received another Rs 100 calling card plus a certificate of participation as motivation.

The course was of 9 days duration from 15 to 23 September 2017. The system was conducted live over the telephony network making use of our GSM gateway which resided in Mauritius. The course was delivered to people of Agalega 1100 Km far from Mauritius over the sea. Nevertheless, our system performed as expected, except that some participants complained about the low sound quality at times.

Figure 2 shows how such a course is conducted over the phone. Users can call a specified number at any time. When connected, an IVR audio interface is presented with an introduction on how to use the system. It then asks them either to start the course from the beginning or to resume the course where they left. The course consists of several lessons based on the specified chapters. Navigation capabilities (pause, play, forward, backward) either through key presses or voice recognition are provided. At any point in time, the user can disconnect. The system registers where the user reached in his course before he disconnects. When the user dials again, the system resumes his course where he stopped.

In addition, the system caters for learner assessment through the yes-no type audio questions. An automatic grading is performed for each learner to know their performance. Students were regularly monitored during the course via the analytics module which provided detailed statistics on the learner behavior on the system, progress achieved, course completion rate, number of lessons done, date and time audio lessons were listened and call duration. Constant reminders were sent over the phone via a pre-recorded audio message to those who lagged behind in their progress.

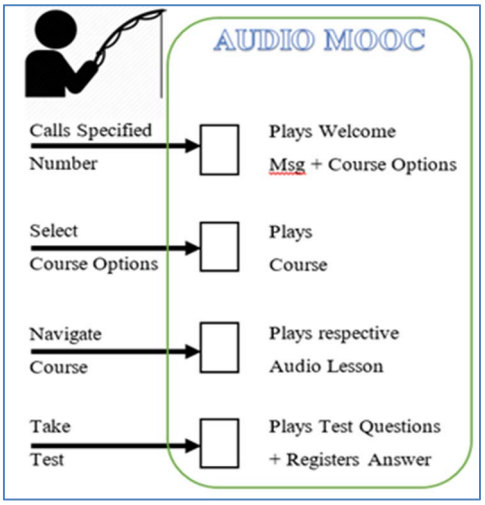


Figure 2. Audio MOOC System

V. RESULTS AND DISCUSSIONS

This section is based on the observations and survey on the 20 Fishermen in Agalega who completed the course in Fisheries Law.

A. System Accessibility

The system was accessible 24/7 without any system downtime recorded. The system performed as expected and was able to deal with usage peaks without any deterioration in terms of quality of service.

B. Demographic attributes of learners and their technology background

Most of the fishermen were above 30 years, out of which only 2 were females. Our investigation showed that the majority of women in Agalega has as main concern their family and household chores with education relegated to a lesser priority. Concerning education and mobile literacy classification, it was noted that 3 youngsters (20-29 yrs) were literate, the rest had very basic literacy education. The majority of them (60%) considered themselves as average in mobile usage knowledge with 10% claiming to be a novice and 30% claiming to be expert. 45% claimed to spend less than 1 hour on call daily, 35 % between 1 and 2 hours while 20% calling for more than 2 hrs. Popular mobile features were music, game, and camera (over 85%) while very few (40%) claimed to know how to use the internet, being very limited and restrictive on mobile data on the island. This gives an insight on their mobile literacy with the ability to use phone features. Hence, irrespective of their literacy status, all of them knew how to operate a phone and make calls.

C. Usability of the System

From the responses provided, the learners used the system without any difficulty and all of them (100%) found that the system interface was very simple and easy to use. This confirms that the system navigation was properly designed since they had no problem in choosing courses from audio menus and navigating through the different lessons. Besides, the system logs denote the pause/ replay/ forward features were used extensively. They also found the assessment conducted over the phone very useful and easy to operate. 100% of the responders found it to be a convenient learning medium since it was accessible anywhere and anytime. This was confirmed on our user analytics module denoting learning activities throughout the day. Besides, they could do other things in addition to listening to the course (80%) which denotes the flexibility of such system. Peaks occurred early morning and after 7 pm, a time where they are free from household chores and work. They found the system good as it is and no improvements were suggested except that at times the sound quality was an issue, but this was the network factor.

D. Course Content

As per the survey, all of the fishermen (100%) were satisfied with the content delivered and found the audio lessons easy to follow. The lessons were recorded in their native language which is Creole, hence eliminating the literacy barriers. The lessons were presented in a formalized and professional way. All of those who completed the course found it beneficial, relevant, worth their time and accommodated their personal learning style at their own pace. The course was successful and brought a change in their behavior (85%). They made requests to have more courses over this medium. This concludes that the assumptions we made in our audio instructional guidelines in terms voice quality, modulation, audibility and duration were good.

E. Course Effectiveness

Conducting a course using the audio-only medium is very challenging since it uses only a single sensory medium as compared to multimedia learning. The Audio MOOC framework attempts to address this issue by proposing a set of audio design guidelines. At the end of the course, we conducted a test on air with the fishermen. A set of 10 yes-no questions were set. It was interesting to note how course completion rate affected performance during the test. Those whose course progress were below 40% could not answer the test questions. Those who completed 60% of the lessons managed to score on average 50%. Participants who successfully followed most of the courses scored above 60%. From the statistical analysis, a strong Pearson correlation coefficient (r=0.8) was obtained showing a positive relationship between course completion and test results. This denotes the effectiveness of the learning medium. The course was well prepared and brought additional knowledge. Hence it was imperative to follow the course to pass.

VI. CONCLUSIONS

This work describes the methodology adopted in conducting a course on Fisheries Law over the phone to 20 fishermen of Agalega Island, a remote island, 1100 Km away from Mauritius. The experiment carried out using the Audio MOOC system has been very conclusive. It allowed us to view our proposed system in action. Since it is an innovative and novel DE medium, it was based on several assumptions in terms of technology, user interface design, pedagogy and instructional design. From the results obtained and the survey conducted, it proved to be a success all across. From a technical and technology perspective, it behaved as expected, easy to manage and deploy courses/tests and track learner behavior. Feedback from participants combined with their test results showed that we were in the right directions in terms of audio pedagogy and instructional design, proving the good foundations of the Audio MOOC framework. The good response we received motivates us to fine-tune the system and conduct more courses in the future. We intend to use the platform for other domains such as agriculture and health. Besides, research on user and learning behavior over such platform still needs to be conducted.

VII. ACKNOWLEDGMENTS

We would like to thank the following collaborators for making Audio MOOC a success: Mr Runden Monique (Research Assistant), Mrs Yajoshi Basant Rai (FiTEC, Mauritius), Mr Vib Senedhun (FiTEC, Mauritius) , Mr Sunil Beeharry (FiTEC, Mauritius) , Mr Ravi Mohit (FiTEC, Mauritius), Mr Neermal Dussooa (FiTEC, Mauritius), Mr Nurungee (OIDC), Mr Naeck (OIDC), Mr Ghurburrun (OIDC) , Mr Kureeman (Resident Manager, Agalega), Mr Patrick Poulay (Assistant Resident Manager, Agalega North), Emmanuel Jasmin (Assistant Resident Manager, Agalega North), Steeve Clarisse(Assistant Resident Manager, Agalega South), Mr Arnaud Meslier (Microsoft, Mauritius).

VIII. SPONSORS

This research is sponsored by the Mauritius Research Council (MRC), Outer Island Development Corporation (OIDC) and Microsoft Mauritius.

REFERENCES

[1] OIDC,2018. Outer Island Development Corporation and Agalega. Online. Available at: http://oidc.govmu.org/English/About%20Agalega/Pages/Infrastructure.aspx

[2] Gov New, 2014. Mauritius and India sign contract for the acquisition of Dornier aircraft. Online. Available at: http://www.govmu.org/English/News/Pages/- Mauritius-and-India-sign-contact-for-acquisition-of-Dornier aircraft.aspx

[3] ICTA, 2018. Beyond IPV4. Online. Available at: https://www.icta.mu/documents/workshop%20ipv6/SESSION2/emtel_IPv6_2011.pdf

[4] Peters, O., 1994. Distance education and industrial production: A comparative interpretation in outline (1973). Otto Peters on distance education: The industrialization of teaching and learning, pp.107-127.

[5] Sherwani, J., Palijo, S., Mirza, S., Ahmed, T., Ali, N. and Rosenfeld, R., 2009. Speech vs. touch-tone: Telephony interfaces for information access by low literate users. In 2009 International Conference on Information and Communication Technologies and Development, ICTD 2009 - Proceedings. p. 447-457.

[6] Agarwal, S.K., Kumar, A., Nanavati, A.A. & Rajput, N., 2010. User-generated content creation and dissemination in rural areas. Information Technologies & International Development, 6(2), pp.pp-21

[7] Chaudry, B.M., Connelly, K.H., Siek, K.A. and Welch, J.L., 2012, January. Mobile interface design for low-literacy populations. In Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium (pp. 91-100). ACM.

[8] Edim, A.E., Muyingi, H. & Sibanda, K., 2013. A Voice User Interface for Low-literacy Users in a Rural Community. International Journal of Computing & ICT Research, 7(1).

[9] Moloo, R.K., 2017. An Audio MOOC framework leveraging the telephony infrastructure as an alternative to Internet Connectivity. Unpublished MPhil Dissertation. University of Mauritius, Reduit, Mauritius.

[10] Shears, A.E., 1984. Development of management courses for the agriculture sector in Nigeria. Programmed Learning, 21(2), pp.88-94.

[11] Long, T., 1984. Broadcasting for rural development. Media in Education and Development, 17(1), pp.17-19.

[12] Ginsburg, M.B. & Arias-Godinez, B., 1984. Nonformal education and social reproduction/transformation: Educational radio in Mexico. Comparative Education Review, 28(1), pp.116-127.

[13] Byram, M. & Kidd, R., 1983. Hands-On Approach to Popularizing Radio Learning Group Campaigns. Convergence, 16(4), p.14

[14] Anzalone S., Bosch A (World Bank) 2005. Improving Educational Quality through Interactive Radio Instruction A Toolkit for Policymakers and Planners. Available at http://siteresources.worldbank.org/INTAFRICA/Resources/toolkit_for_radio.pdf

[15] Cooksun, P.S. (1997): Instructor & Participant responses to critical condition of Audio conferencing. Tech. Trends 40 (4): pp. 23-25

[16] Macmullen, P., 2001. Audio conferencing in support of distance education. COL Available at http://oasis.col.org/bitstream/handle/11599/34/KS2001-02_audioconferencing.pdf?sequence=1&isAllowed=y

[18] Gilroy K., Diedhiou A., Cox C.M., Duncan L., Koumtingue D., Pacqué-Margolis S., Fort A., Settle D., Bailey R., 2015. Use of an Interactive Voice Response System to Deliver Refresher Training in Senegal: Findings from Pilot Implementation and Assessment. Available at http://www.capacityplus.org/files/resources/ivr-system-refresher-training-senegal.pdf

[17] Hitch, L.P. & MacBrayne, P., 2003. A model for effectively supporting e-learning. [Online]. Available at: http://technologysource.org/article/6/

[19] Vuong, S.T., Schroeder, J., Alam, M., Tjia, A., Chung, Y.I. and Chen, A., 2010. LIVES: Learning through an interactive voice educational system. In Proceedings of the IADIS International Conference Mobile Learning 2010. pp. 157-164. [Online] Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84860738462&partnerID=tZOtx3y1.

[20] Motiwalla, L.F., 2009. A voice-enabled E-learning service (V??IS) platform. in: Proceedings of the 5th International Conference on Networking and Services, ICNS 2009. pp. 597–602.

[21] García, V.M.Á., Ruiz, M.D.P.P. & Pérez, J.R.P., 2010. Voice interactive classroom, a service-oriented software architecture for speech-enabled learning. Journal of network and computer applications, 33(5), pp.603-610

[22] Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 63, 81-97

[23] Gronlund, N.E., 1998. Assessment of student achievement. Allyn & Bacon Publishing, Longwood Division, 160 Gould Street, Needham Heights, MA 02194-2310; tele.

[24] Guo, P.J., Kim, J. and Rubin, R., 2014. How video production affects student engagement: An empirical study of mooc videos. In Proceedings of the first ACM conference on Learning@ scale conference (pp. 41-50). ACM.

[25] Keller, J.M., (1987). Development and use of the ARCS model of instructional design. Journal of instructional development, 10(3), pp.2-10.

[26] Bloom, B.S., 1956. Taxonomy of educational objectives: The classification of educational goals.

[27] Bellis, T 2003, Assessment and management of central auditory processing disorders in the educational setting: from science to practice, Thomson Delmar Learning, Canada.

Enhanced Search for Arabic Language Using Latent Semantic Indexing (LSI)

Fawaz S. Al-Anzi
Department of Computer Engineering
Kuwait University
Kuwait City, Kuwait
fawaz.alanzi@ku.edu.kw

Dia AbuZeina
Department of Information Technology
Palestine Polytechnic University
Hebron, Palestine
abuzeina@ppu.edu

***Abstract*—However, this technique poses some challenges such as high dimensional space and semantic loss representation. Therefore, the latent semantic indexing (LSI) is proposed to reduce the feature dimensions and to generate semantic rich features that represent conceptual term-document associations. In particular, LSI has been successfully implemented in search engines and text classification tasks. In this paper, we propose a novel approach to enhance the quality of the retrieved documents in search engines for Arabic language. That is, we propose to use a new extension of the LSI technique instead of just using the standard LSI technique. The LSI method proposed is based on employing the word co-occurrences to form a term-by-document matrix. The proposed method is to be based on the documents evaluating cosine similarity measures for term-by-document matrix. We will empirically evaluate the performance using an Arabic data collection that contains no less than 500 documents with no less than 30,000 unique words. A testing set contains keywords from a specific domain will be used to evaluate the quality of the top 20-30 retrieved documents using different singular values (i.e. different number of dimensions). The results will be judged on the performance of the proposed method as it is compared to the standard LSI.**

Keywords—Arabic text, latent semantic indexing, searche engine, dimensionality reduction

I. INTRODUCTION

Due to the explosive growth of online information, search engines have a prominent role in information retrieval (IR) and web mining applications. The web is the largest repository of public data that undoubtedly requires efficient algorithms for retrieving and filtering out the textual information as well as other object types. Hence, search engines are becoming more intelligent in filtering desired content. In general, textual data is represented using the Vector Space Model (VSM) where each document is represented using a vector of components (i.e. attributes), many of which could be zero. However, VSM poses some challenges such as there are lengthy feature vectors and semantic loss representation. Therefore, the Latent Semantic Indexing (LSI) technique is proposed to alleviate such challenges and hopefully to enhance the performance. LSI aims at transforming the original textual vectors into conceptual vectors that are characterized by two distinguished properties: reduced dimensions and semantic rich features. The intrinsic nature that determines the quality of the LSI is found in the semantic property that enforces returning semantically close documents without the constraint to have the exact searching keywords.

LSI is based on a theorem from linear algebra that is called Singular Valued Decomposition (SVD). The SVD can transform the textual data, which is represented as a large term-by-document matrix into a smaller semantic space represented as three matrices where the product of the generated matrices equals the original term-by-document matrix. Hence, the first step of LSI is to decompose (i.e factorize) the term-by-document (A) matrix as follows: A=USVT where U is a matrix that provides the weights of terms, S provides the eigenvalues (also called singular values) for each principal component direction, and VT is a matrix that provides the weights of documents. VT is the matrix that contains the document feature vectors that are generally used in IR and text mining applications. In the standard implementation of LSI starts with a term-by-document matrix to generate the required feature vectors that are used for classification. However, a term-by-document matrix generally is formed using different values such as (Boolean flags, counts, or weights) to track the occurrences of terms in documents. For classification, the generated LSI feature vectors are generally utilized using a similar measure such as Euclidian, Mahalanobis, Manhattan, cosine similarity, etc. The cosine similarity measure is known to be one of the popular distance measures in pattern recognition. In this paper, we propose an extension of the LSI implementation by using the cosine measures instead of the standard word co-occurrences values. Hence, the proposed method forms the term-by-document matrix using the cosine measures between documents before employing the SVD process.

In the next section, we present the motivation followed by the literature review in section 3. In section 4, we present the proposed method followed by the experimental results in section 5. We conclude in section 6.

II. MOTIVATION

Text mining systems are intuitively in need for very efficient algorithms that intelligently understand the search engine's documents as well as the query's keywords. Moreover, huge online data requires considering the semantic relationships between both the documents and the words that are also called co-occurrences. Unlike trivial searching methods that are based on traditional text matching, LSI is characterized by semantic rich value that enable the system to return useful results without having exact matching words between the document and the query's keywords. For example, if we search for the word "coffee," it is expected that the system will return many documents related to this word, however, it might return other related documents that have no "coffee" word in it, but are semantically related to the word "coffee." That is, it is possible to obtain documents that belong to the topic, such as stimulant effects, caffeine, etc. Using the enhancing searching process with thorough overwhelming digital data requires an endless research effort to satisfy the users'

requests. In fact, text mining is a challenging task since documents usually have mixed contents that make it difficult to digitally understand the document's category. For an illustration, Fig. 2 shows a document that has different words, such as "headaches": "صداع", "addiction": "ادمان", "drugs": "مخدرات", "tension": "التوتر", "frenzy": "الهيجان", and "crimes": "الجرائم". Such diverse words make it vague for IR algorithms to search correctly for the required data. By nature, medical documents require precise algorithms that can adequately find the proper documents for the user.

The LSI has been proven to be a valuable tool that reveals the semantical relationship between data objects (i.e: the words in this research). Based on detected underlying semantic distinctions, LSI is able to bring out the relevant documents that do not contain the searching keyword at all.
Literature Review

In the literature, there are many studies that discuss the LSI technique. In particular, LSI is used for the text mining task, such as text classification, text summarization, text clustering, search engines, etc. LSI initially was presented by Deerwester in [1] as a standard dimension reduction technique in IR. Reference [2] presents an algorithm to enhance the results of search engines. The algorithm combines common phrase discovery and LSI techniques to separate search results into meaningful groups. Reference [3] presents a new implementation of the standard LSI. The new implementation aims to provide efficient, extensible, portable, and maintainable LSI. Reference [4] presents a theoretical model for understanding the performance of LSI in retrieval applications. Reference [5] presents an LSI based method for fully automated cross-language document retrieval in which no query translation is required.

Reference [6] describes a word clustering approach that is based on LSI. Reference [7] proposes a local LSI method called "Local Relevancy Weighted LSI" to improve text classification by performing a separate SVD on the transformed local region of each class. Reference [8] uses LSI to automatically identify the conceptual gene relationships from titles and abstracts in a database citation. Reference [9] proposes and empirically tests the feasibility and utility of post-retrieval clustering of digital forensic text string search results – specifically by using Kohonen Self-Organizing Maps (SOM) as a self-organizing neural network approach. Reference [10] proposes a hybrid term frequency – inverse document frequency (TF-IDF) that is based on algorithm and a clustering based algorithm for obtaining multi-post summaries of Twitter posts along with the detailed analysis of Twitter post domain. Reference [11] uses LSI for automatic software clustering. LSI was used as the basis to cluster software components, source code, and its accompanying documentation. Reference [12] proposes two text summarization approaches: the modified corpus-based approach (MCBA) and the LSI-based approach.

LSI has been widely documented as a retrieval method that employs SVD for semantic rich reduced feature vectors. Nevertheless, utilizing LSI and SVD requires understanding which values in the reduced dimensional space contain the word relationships (latent semantic) information. Hence, many studies in the literature have discussed this important aspect. Reference [13] presents an empirical study of the required dimensionality for large-scale LSI applications. Reference [14] was developed as a model for understanding which values in the reduced dimensional space contain the term relationship (latent semantic) information.

Regarding cosine similarity, it is a well-known similarity measure that has been widely mentioned in the literature. Reference [15] indicates that cosine similarity dominants have similar measures in IR and text classification. This measure is based on the cosine of the angle between two vectors. Reference [16] demonstrates that the similarity between two documents can be measured using the cosine of the angle between the two document feature vectors, which are represented by using VSM. Theodoridis and Koutroumbas in [17] defines the cosine similarity measure as: $Scosine(x,y) = xTy / \| x \| \| y \|$ where $\| x \|$ and $\| y \|$ are the lengths of the vectors x and y, respectively. Reference [18] proposes that the cosine similarity is a robust metric for scoring the similarity between two strings. Reference [19] demonstrates that the cosine similarity is used to find the vectors neighborhood. Reference [20] demonstrates that the cosine similarity is easy to interpret and simple to compute for sparse vectors, this indicates that it is widely used in text mining and IR. Reference [21] used the cosine similarity measure for the Arabic language text summarization. Reference [22] uses the cosine measure for the language identification problem.

III. The Proposed Method

This work proposes an enhanced method for the standard LSI. Initially, the preprocessing step was performed by declaring the stop words and ignoring the characters' list. In addition, all small words that were less than three characters in length were discarded. A normalization process was also performed to change some Arabic characters such as (إ→ا) and (أ→ا). As shown in Fig. 1, the term-by-document (A) matrix was created using the unique words in the used corpus. The term-by-document (A) matrix weighted used TF-IDF. For comparison purposes between the proposed method and the standard LSI, A was decomposed into three matrices (U: Term by dimension; S: Singular values; and VT: Document by dimension). The diagonal of matrix S contains singular values so that one can choose the desired reduced dimensions. In general, not all singular values were considered; instead, only the most important values were considered starting from the first singular values up to the desired value (k).

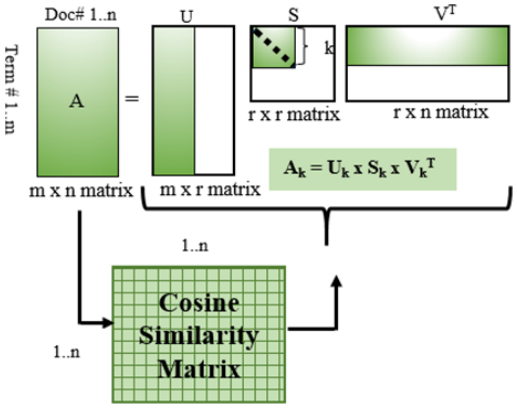


Fig. 1. Forming cosine similarities matrix after SVD.

The proposed method has an extension of the standard LSI by creating a new matrix called the cosine similarity

matrix. This new matrix has used the cosine similarities between all documents in the corpus instead of the co-occurrences (i.e. instead of the frequency of a word in a document) that usually are used when creating term-by-document matrices. Hence, the enhanced method is summarized by using four main steps as follows: 1) creating the standard term-by-document matrix using word co-occurrences; 2) the matrix is weighted using TF-IDF; 3) based on the standard term-by-document matrix, we formed a new matrix called cosine similarity matrix that contains the cosine measures between each two vectors in the standard term-by-document matrix ; and then finally, 4) the SVD is used to truncate the cosine similarity matrix to generate the enhanced feature vectors that are used in the search engine. Of course, different singular values (k) might be investigated to find the optimal performance. As our corpus contains 800 documents, the cosine similarity matrix of the used corpus is of the size 800 × 800. Of course, the cosine similarity matrix is a symmetrical matrix.

In both cases, the standard LSI or the proposed method, the query's keywords have to transfer to the LSI space. For the standard LSI, the query's feature vectors transform into the new reduced space that is called "folding-in." This is done by using the following formula: V′ T=AUS-1. Hence, V′ T contains the reduced query's feature vectors that are used along with VT in the classification process. For the proposed method, the query's feature vectors have two transformation steps. The first is regarding the cosine measures against all feature vectors of the training documents before using the folding-in technique as a second step (i.e. similar to the standard LSI but for the cosine measure instead of the word co-occurrences). Fig. 2 shows how to generate the query's vector in terms of cosine similarity. Hence, the cosine similarity matrices of the training and the testing set are generated for the new SVD implementation.

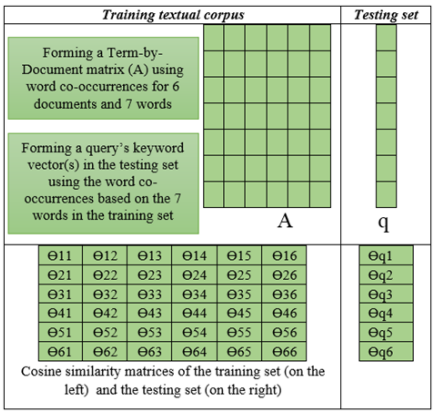


Fig. 2. An example of creating cosine similarity matrices.

IV. The Experimental Results

The proposed method was evaluated using an Arabic textual corpus containing 800 documents, 353,888 words, and 47,222 unique words. The data collection was in regards to medical stories obtained from Alqabas [23], a Kuwaiti newspaper. A testing set contained five medical keywords that were used as queries for the developed search engine. Since the number of singular values is important in LSI applications, we considered a wide range of singular values to measure the performance for both cases (i.e. standard LSI and the proposed method). Hence, the search engine was evaluated using the different number of feature vectors dimensions (k). That is, a series of experiments were performed using the following k :{ k=10, 20, 30, 40, 50, 60, 70, 80, 90, 100,150,200, 250,300, 350,400, 500}. At each singular value, we analyzed the top-20 retrieved documents to investigate the query's keyword occurrences.

Table 1 shows the performance of the word "الزهايمر" : "Alzheimer." In the table, the first row indicates the medical query keyword that is the first word in the testing set. The first column indicates the singular values k that starts at 10 and ends at 500. At k=10, the word "الزهايمر" : "Alzheimer" is found in the first document zero times while it was found in the top-20 retrieved documents three times, using the standard LSI as shown. On the other hand, it was found one time in the first document and 4 times in the top-20 retrieved documents using the proposed method. The results show that the retrieved document using the proposed method is of a high quality compared to the standard LSI even with lower dimensions. For an illustration, at k=80, the standard LSI retrieved a document that contained 1 occurrence of the searched word with 11 occurrences in the top-20 documents, while the proposed method returned a document that contained 6 occurrences with 20 occurrences in the top-20 documents. Table 2 also shows that the maximum occurrences of the word "الزهايمر" : "Alzheimer" is 27 times using standard LSI, however, it scored 29 occurrences using the proposed method. The results presented in Table 2 did not require the exact match cases as we considered the word "الزهايمر" : "Alzheimer" to be the same as saying "the Alzheimer" "زهايمر" and "بالزهايمر ", etc. Hence, different variations of the same word were counted.

TABLE I. Searching results for different dimensions of "Alzheimer"

"الزهايمر : "Alzheimer"				
	The Standard LSI		The Proposed Method	
k	First doc.	Top-20 doc.	First doc.	Top-20 doc.
10	0	3	1	4
20	0	6	6	13
30	0	7	6	13
40	0	10	6	18
50	1	3	6	18
60	1	7	6	18
70	1	11	6	18
80	1	11	6	20
90	1	7	6	16
100	1	11	6	16
150	1	19	6	21
200	1	21	6	23
250	1	18	6	23
300	1	18	6	25
350	3	24	6	25
400	3	27 (max)	6	26
500	3	25	6	29 (max)

In addition, we evaluated the performance by measuring the percentage of the matched words among all occurrences in the training set. For example, the word "اشعة" : "rays" appears 324 times in the corpus. The standard LSI showed this word 112 times in the top-20 list as indicated in table 6. However, the proposed method listed it 116 times. Hence, the percentage for the standard LSI is 112/324=0.346. For the proposed method, the percentage is 116/324=0.358. These percentages are shown in Table 2 for all words of the testing set. The table also shows that the average of the percentages for the standard LSI is 0.571 and for the proposed method is 0.629. This means that the proposed method outperforms the standard LSI by 5.83% for the top-20 retrieved documents.

TABLE II. The percentage of the retrieved searching words

#	Word	The Standard LSI	The Proposed Method
1	"الزهايمر" : "Alzheimer"	0.844	0.906
2	"فيروس" : "virus"	0.520	0.559
3	"الاوكسجين" : "oxygen"	0.309	0.473
4	"القهوة" : "coffee"	0.839	0.853
5	"اشعة" : "rays"	0.346	0.358
	Average	0.571	0.629

Other evaluation methods should be required since we only compared the match word while the semantic quality of the retrieved documents should also be evaluated. Fig. 3 shows the graphical representation of the performance differences between the standard LSI and the proposed method. The graph's information is based on the percentages calculated in Table 7.

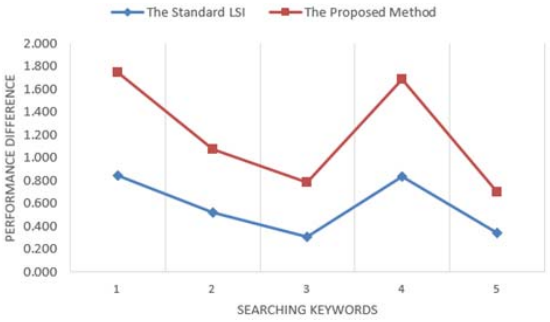


Fig. 3. The performance enhancement using the proposed method.

Finally, the proposed method is suitable for relatively small data collections. However, it might not be very efficient for very large corpora that contains millions of documents. In fact, creating the cosine similarity matrix O(n2) where n is the total number of documents in the corpus is very complex and requires an extensive amount of time. Nevertheless, this method shows a possible enhancement especially when we look for precise results for highly mixed contents as medical documents. In addition, the recently available high speed machines might help to improve query searches over time with greater complexity. Moreover, as quickly as technology is evolving, in the future, we will be able to utilize these advancements with greater accuracy, faster speeds, and more enhanced data.

V. Conclusion

This paper presented a new variant of the LSI technique for search engines. A comprehensive experimental evaluation shows the feasibility of the LSI technique as well as the enhancements of the new method over the standard LSI technique. The results showed that using the documents' cosine similarities instead of just word co-occurrences enhances the performance of search engines. The proposed method shows that the top-20 retrieved documents are of a higher quality than the top-20 documents retrieved using the standard LSI. As a future work, we propose to investigate the proposed method for larger data collections as well as investigating the time and space complexities of the proposed method. In addition, the evaluation should include the semantic quality and not just matched words.

Acknowledgment

This work is supported by Kuwait University Research Administration Research Project Number EO03/18.

References

[1] Deerwester, Scott, Dumais, Susan T., Furnas, George W., Landauer, Thomas K., and Harshman, Richard. "Indexing by latent semantic analysis." Journal of the American society for information science 41.6 (1990): 391.

[2] Osiński, Stanislaw, and Dawid Weiss. "A concept-driven algorithm for clustering search results." Intelligent Systems, IEEE 20.3 (2005): 48-54.

[3] Letsche, Todd A., and Michael W. Berry. "Large-scale information retrieval with latent semantic indexing." *Information sciences* 100.1 (1997): 105-137.

[4] Kontostathis, April, and William M. Pottenger. "A framework for understanding Latent Semantic Indexing (LSI) performance." Information Processing & Management 42.1 (2006): 56-73.

[5] Dumais, Susan T., Littman, Michael L. and Landauer, Thomas K.. "Automatic cross-language retrieval using latent semantic indexing." *AAAI spring symposium on cross-language text and speech retrieval*. Vol. 15. 1997.

[6] Bellegarda, J. R., Butzberger, J. W., Chow, Y. L., Coccaro, N. B., & Naik, D. (1996, May). A novel word clustering algorithm based on latent semantic analysis. In Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on (Vol. 1, pp. 172-175). IEEE.

[7] Liu, T., Chen, Z., Zhang, B., Ma, W. Y., & Wu, G. (2004, November). Improving text classification using local latent semantic indexing. In Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on (pp. 162-169). IEEE.

[8] Homayouni, R., Heinrich, K., Wei, L., & Berry, M. W. (2005). Gene clustering by latent semantic indexing of MEDLINE abstracts. Bioinformatics, 21(1), 104-115.

[9] Beebe, Nicole Lang, and Jan Guynes Clark. "Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results." Digital investigation 4 (2007): 49-54.

[10] Inouye, David, and Jugal K. Kalita. "Comparing twitter summarization algorithms for multiple post summaries." Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on. IEEE, 2011.

[11] Maletic, Jonathan, and Naveen Valluri. "Automatic software clustering via latent semantic analysis." Automated Software Engineering, 1999. 14th IEEE International Conference on.. IEEE, 1999.

[12] Yeh, J. Y., Ke, H. R., Yang, W. P., & Meng, I. H. (2005). Text summarization using a trainable summarizer and latent semantic analysis. Information processing & management, 41(1), 75-95.

[13] Bradford, Roger B. "An empirical study of required dimensionality for large-scale latent semantic indexing applications." Proceedings of the 17th ACM conference on Information and knowledge management. ACM, 2008.

[14] Kontostathis, April. "Essential dimensions of latent semantic indexing (lsi)."System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on. IEEE, 2007.

[15] Elberrichi, Z., Rahmoun, A., & Bentaallah, M. A. (2008). Using WordNet for Text Categorization. Int. Arab J. Inf. Technol., 5(1), 16-24.

[16] Beil, Florian, Martin Ester, and Xiaowei Xu. "Frequent term-based text clustering." Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002.

[17] Theodoridis, S. and K. Koutroumbas (2008). Pattern Recognition, Fourth Edition, Academic Press.

[18] Tata, S., & Patel, J. M. (2007). Estimating the selectivity of tf-idf based cosine similarity predicates. ACM Sigmod Record, 36(2), 7-12.

[19] Chattamvelli, R. Data Mining Algorithms, Published by Alpha Science International Ltd., 2011

[20] Dhillon, I. S., & Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. Machine learning, 42(1-2), 143-175.

[21] Sobh, I., Darwish, N., & Fayek, M. (2006). A trainable Arabic Bayesian extractive generic text summarizer. In Proceedings of the Sixth Conference on Language Engineering ESLEC (pp. 49-154), Ain Shams University, Cairo, Egypt.

[22] Takçı, H., & Güngör, T. (2012). A high performance centroid-based classification approach for language identification. Pattern Recognition Letters,33(16), 2077-2084.

[23] Alqabas. (2016, October). Retrieved from http://www.alqabas.com.kw/Default.aspx.

Scaling Education in Emerging Markets to Participate in Industry 4.0

Johannes Wynand Lambrechts, *Senior Member, IEEE*
Department of Electrical and Electronic Engineering Science,
Faculty of Engineering and the Built Environment
University of Johannesburg
Johannesburg, South Africa
wynand.lambrechts@ieee.org

Saurabh Sinha, *Senior Member, IEEE*
Deputy Vice-Chancellor, Research and Internationalisation
University of Johannesburg
Johannesburg, South Africa
dvc-r@uj.ac.za

***Abstract*—A modern bottom-up approach to decentralizing education through information and communication technology (ICT) can have distinct advantages in expanding, replicating and adapting (scaling) education. Learners benefit from this approach through garnering not only knowledge and information, but also skills development in preparation for the fourth industrial revolution (Industry 4.0). However, the challenges faced by emerging markets (interchangeably referred to as developing countries) are vastly different when compared to developed countries. Adoption and progress are significantly slower in these regions. A decentralized education system is a precursor for a decentralized industry, one of the primary characteristics of Industry 4.0. The unique challenges in the developing world are often overlooked and as a result socioeconomic growth is stifled and sustainability is threatened. In this research, the challenges and limitations that emerging markets are experiencing are identified, and a specific approach to scaling education through ICT is reviewed. The advantages of scaled and decentralized education are also acknowledged, and the re-evaluation of policies governing learners and educators in developing countries presented.**

Keywords—ICT in education, Industry 4.0, decentralization, scaled education, sustainability, deconcentration, developing countries, emerging markets, education, pedagogy

I. INTRODUCTION

Innovation (evolutionarily and revolutionary) should be a top priority in the higher education system of any country [1]. Central to modern innovation, economic growth and long-term sustainability is technology. In many countries, especially in developing countries, there are inconsistencies between the education pursued by students and qualifications that inspire learners to lead successful start-up businesses and innovate new career paths in line with the fourth industrial revolution (Industry 4.0). This shortfall is intensified by a lack of clarity of the direction of Industry 4.0 and rapid changes in the requirements of the labor market. The heightened use of information and communication technology (ICT) to facilitate electronic learning (e-learning [2]) requires widespread retraining of educators and changing what information is transferred to learners and how this happens. The extent and pace at which industry can exploit ICT in sectors such as manufacturing are governed by existing infrastructure, local ICT adoption and worker skill sets [3].

A. Research objectives

In developing countries, digital education and vocational training are not receiving governmental support at the levels required to prepare these countries to take part in a paradigm shift affecting how business is conducted globally. Industry 4.0 thrives on a decentralized economy and offers a means for countries to maintain competitiveness. A decentralized education system that is opportunistically scaled rather than spontaneously diffused share the advantages and limitations of decentralizing an economy, but ultimately enables financially incapable and geographically dispersed students to learn the skills essential in an evolving world.

In this paper, the problems of slow and lackluster scaling of education in emerging markets have been identified. The authors aim not only to identify the challenges, but also to provide potential solutions that could mitigate challenges through sustainable methodologies. As part of an emerging market population, understanding South Africa's historic past and being aware of the long-term effects of sub-standard education, the authors aim to attract attention from local and international organizations capable of providing partnered support based on the findings in this paper.

B. Research methodology

For countries and individuals to participate in Industry 4.0, education should be scaled to facilitate learning and teaching techniques that complement sustainability and socioeconomic development, particularly in developing countries [4]. Participating in Industry 4.0 entails actively engaging, collaborating, innovating and ultimately creating sustainable and long-term socioeconomic wellbeing for citizens through the opportunities that technology presents. Quantifying the success of a policy aimed at achieving this is often difficult to achieve and the effects are only seen over a long period (sometimes only after a generation). This work identifies techniques that have proven to be successful, typically in developed countries, and isolates additional challenges apparent in emerging markets, including financial, political and socioeconomic limitations. Furthermore, technology is emphasized as the primary enabler in achieving scaled education in emerging markets.

C. Scope of the paper

The following section describes the primary characteristics of a decentralized educational system, as well as a decentralized industry. The reviews serve as a precursor to the subsequent section, identifying the challenges of scaling education in emerging markets, and technology's role in modern education and decentralization. In the final section, the paper introduces techniques that re-evaluate education, specifically in emerging markets, and can potentially benefit the process of scaling education in the developing world.

II. An Evolution in Scaling Educating

A. Centralized and decentralized education

Centralized education historically emerged from early apprenticeship-style teachings where isolated communities pursued their own objectives [3]. In many developing countries, the traditional top-down approach to schooling, content mastery, is still being implemented. Although an effective approach in numerous circumstances, it can still result in squandered resources and foregone opportunities if not combined with a bottom-up approach and purposefully scaled. Education is not only measured by cognitive development. According to the American Psychological Association, socioeconomic status (SES) is not only an embodiment of the gross income of individuals, but also a measure of their financial security and educational attainment and a generalized perception of social status and class. In low SES communities, several identifying factors contributed by underdeveloped education include:

- difficulty in storing and remembering knowledge, information, and other cognitive skills,
- poor spoken and written language development,
- sub-standard psychological development, and
- inability to process socioemotional relationships.

Rapid population growth, urbanization, nationalism, diversification and the internet have all played their part in evolving the role of scaling education in a globalized world. Education systems altogether are following a trend towards different forms of decentralization, either functional or territorial. In a centralized system, decision-making is regulated solely by the education ministry, which controls all facets of the structure, including factors related to learners, educators and educational facilities [5]. Through decentralization, decision-making (and accountability) are transferred to local and intermediate governments, communities and educational institutions, with limited regulation granted to the central government. Fig. 1 is a simplified representation of a centralized versus a decentralized system, applicable in not only education.

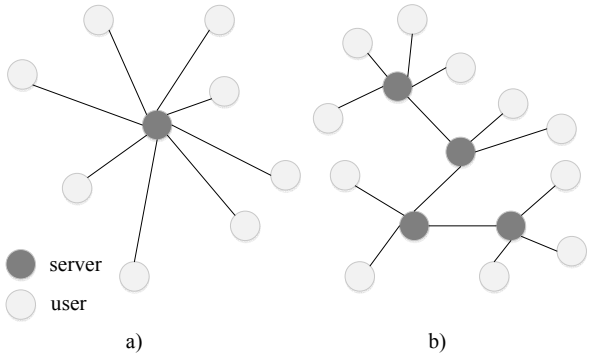


Fig. 1. A simplified representation of a a) centralized versus a b) decentralized system

The extent of decentralized transfer of decision-making varies from administrative to broader affairs of financial control and curriculum deconcentration. The theoretical advantages of decentralized education, such as transparent decision-making, administrative and fiscal efficiency and access to high-quality information and services, are appealing, especially when considering the severe lack thereof in developing countries. In particular, governments with extensive fiscal constrictions are attracted by the potential of decentralization to proliferate productivity, develop the skill sets of the population and increase socioeconomic growth.

B. Centralized and decentralized industry

Sustainability and economic growth can thrive if inclusive development throughout the working population is stimulated [6]. Industry 4.0 is merging with socioeconomic and demographic elements to spearhead momentous changes in business models across numerous activities. A disruption in the labor force is introducing new classes of professions and careers, often displacing or altering well-known occupations. Traditionally, a centralized manufacturing organization had one principal facility to manufacture and distribute goods or services, or a principal site with numerous distribution points within the supply chain. A centralized site can reduce per-unit production cost by sharing equipment in manufacturing processes for different products and achieving economies of scale. Centralized manufacturing is also traditionally associated with creating more local jobs, consistent production and effective use of resources. However, in emerging markets, especially in rural areas, people are often unable to afford travelling costs to and from these factories. Evolving technologies, sociodemographic shifts and political and economic uncertainties in developed and emerging markets are already reshaping the workforce. The traditional sense of an employee is evolving with Industry 4.0 and early adopters of this change are already reaping the benefits. The employee of yesteryear was

- geographically confined,
- expected to work during agreed-upon office hours,
- focused on knowledge and on inputs, and
- restricted to corporate learning and teaching.

A decentralized industry and digital economy requires an employee with significantly transformed characteristics, an adaptive education, exceptional problem-solving skills and critical thinking. The internet enables employees to work from anywhere, at any time, and to focus on outputs that enhance service offerings, efficiency and the products of the industry. Furthermore, a decentralized industry empowers employees and improves the efficiency of decision-making by focusing on challenges and issues experienced by individual employees, rather than focusing on *one-size-fits-all* solutions. While decentralized industries have numerous advantages for employees, investments in multiple sites, company-wide consistency and transferring materials between sites have higher cost implications for the organization and investors. In emerging markets, these risks often outweigh the benefits of a decentralized approach.

III. Scaling of Education in Emerging Markets

In many developing countries, decentralizing of education as a form of maintaining quality has limited deconcentration of functions to local communities financing their own schools [7]. Unfortunately, in emerging markets,

successful scaling and decentralization are often not a result of successful government policies, but rather a result of failure to deliver basic services [7]. In such circumstances, lack of access to education for learners requires commitment from the community to implement tailored policies that deliver quality knowledge transfer to learners dedicated to educating themselves. Developing countries typically have a relatively high success rate with informal or formal parental participation in schooling, but governmental mechanisms needed to assign roles and responsibilities as well as accept accountability are poor. International organizations such as the World Bank have intervened on several occasions to attempt to fortify the ability of governments to manage education and scale systems, often resulting in administration of resident infrastructure projects. In the modern era, Industry 4.0-focused decentralization is essential and scaling organization of instruction, personnel management, planning of curricula and allocation of resources are focused on a digital age of education.

A. Unique challenges faced by emerging markets

Garnering support from government, public and private corporations, communities and individuals to develop and maintain policies structured to provide people with education indicative of a global approach is a huge task, and requires devoted patronage from all parties involved. Several international and multilateral initiatives have been developed to support poor communities to achieve bottom-up combined with top-down universal education, including

- the United Nations prioritizing universal primary education as one of the Sustainable Development Goals (SDGs) for 2030,
- the Fast track Initiative as a partnership between supporting countries and emerging markets to speed up progress towards the SDG,
- the private sector, specifically the World Bank and the United States Agency for International Development, promoting public-private partnerships to contribute to global education, and
- various bilateral agreements dedicating financial assistance to poor countries to increase access to educational institutions through teaching and re-learning educators, providing educational supplies such as textbooks and incorporating technology as a learning tool.

In developing countries, however, there are numerous additional challenges that are often prioritized over providing citizens with quality education. A disconnected population, with no access to the internet, is often overlooked as a means to grow and sustain socioeconomic development over the long term. Electronic commerce (e-commerce), big data, and social media are contributing to how business is conducted globally, but also require knowledge of the ICT sector, adaptive education, and most importantly, connectivity. Socioeconomic stability will inevitably follow; however, this is a long process and developing countries are facing more pressing issues that are frequently not considered in the developed world. Specific challenges to achieving a growing education system and empower learners with conceptual knowledge through connectivity include

- lack of infrastructure that is needed to facilitate policies,
- physical lack of infrastructure in the ICT sector, energy supply, and the education sector,
- lack of skilled workers and undertrained educators,
- underdeveloped financial markets and lack of affordable quality education,
- corruption reaching the highest levels of the state, trickling down to the ICT and education sectors, and
- epidemics in public health, poverty, hunger, and high levels of inequality.

The United Nations SDGs for 2030 acknowledge quality education, good jobs and economic growth, sustainable cities and communities, and innovation and infrastructure as global challenges, with emphasis on the developing world. It remains a momentous challenge to implement these goals and garner support from local governments.

Dramatic changes are needed where governments and non-governmental organizations that invest in education in developing countries tend to quantify the success rate based on test scores [8] and student/staff feedback. As the modern world is changing from its focus on knowledge to a focus on adaptive learning, these models also need to be adapted. Supporting organizations and governments need to look beyond these traditional models and develop knowledge, skills, and attitudes that will have the greatest social and economic impact on people living in developing countries.

B. Technology's role in education

The Merriam-Webster dictionary defines pedagogy as the art, science, or profession of teaching. Technology should not be integrated into pedagogy just for the sake of using technology. Such a modus operandi is unlikely to lead to long-lasting or sustainable socioeconomic wellbeing. Technology should act as a supporting structure in the educational process, whereas scaling is a direct result of using technology. The use of technology as an educational aid must have a clearly defined objective, guiding policies and strategies to facilitate its implementation and a maintainable model. The integration of ICT into education already has multiple well-defined frameworks that define technology-focused models, such as Bloom's taxonomy, the TPACK framework, NIMB frameworks, and the UNESCO framework [9]. Central to all these models is a thematic approach to describe the progression of learning and teaching. Identifying the role that ICT plays in each of these frameworks is essential to support the modern progression of scaling and/or decentralizing learning and teaching. In developing countries, additional considerations of fiscal constraints, a limited number of educated citizens to drive the frameworks, and often geographically remote rural areas all present unique and additional challenges. As outlined in [9], although current strategies and policies in certain developing countries (such as South Africa) exist, implementation is slow and fiscal and human resources are limited. A definite shift from learner performance to capacity development (knowledge-based towards adaptive skills-based) is required as ideally, a phased approach with suitable

short-term objectives must be adopted. Several secondary advantages for learners that follow suit when uniting ICT and education include

- increased encouragement to learn and engagement,
- encouraging active participation leading to higher levels of knowledge retention,
- tailored education to inspire and encourage individual learning and/or collaboration, and
- developing useful secondary skills that are required by Industry 4.0 criteria.

There are of course certain disadvantages of using ICT in education. The use of ICT as an educational tool should be combined with, ironically, educating learners and educators in using it effectively. Common disadvantages include that

- skilled staff are required to maintain the infrastructure and the hardware/software,
- it is expensive to develop the infrastructure, especially in developing countries where current infrastructure is severely lacking or non-existent,
- ICT services can be misused, act as a distraction, or misguided information can be accessed (fake news), and
- learners can become disconnected from the real world and have under-developed dexterity and social skills and even develop health issues from extended periods of physical inactivity.

Again, many of these disadvantages or challenges that arise from ICT in education are often amplified in developing countries, especially considering the financial implications of a digitally integrated infrastructure. Furthermore, increased dependence on a reliable electricity supply when using ICT as a tool is problematic in many developing countries.

C. Technology's role in modern decentralization

Decentralization essentially gives an individual more ability to effect changes within a system and is a useful tool to achieve scaling of various sectors, including education. It is empowering and already changing fundamental financial systems, innovative new technologies (block chain and cryptocurrencies are based on decentralization), and even political structures. Through a digitized decentralized approach, entrepreneurial individuals are facilitated to transform ideas into products (3D printing) or services, using a collection of technologies, with the internet at the core in numerous scenarios.

The work in [10] investigates how the allocation of authority within organizations (of all sizes) is changing as the industry is adapting to an evolving information structure. It is predicted that firms that rely heavily on technology will be more likely to opt for a combination of a centralized and decentralized approach, primarily because new technologies inevitably mean that there is limited information within the current body of knowledge. Similar predictions are proposed in [10] with regard to organizations in heterogeneous environments as well as start-up/young firms, which are all likely to adopt a decentralized approach to gather new information. Older, established firms are more likely to have a centralized model in place, relying on past experience and knowledge garnered in the organization to drive future products and services.

An example of advances in technology leading to a decentralized and ultimately scaled system is in the path to energy independence. Especially relevant to developing countries and rural areas, energy independence is crucial when considering a technology-driven decentralized (education) system. Referring to Fig. 1 a, energy production in most parts of the industrial world is a centralized process. Utility companies invest large capital in infrastructure, maintenance and delivery of energy to homes and businesses. As a result, all individuals are reliant on the utility for their energy requirements. Inefficient, inconsistent or unreliable energy generation by the utility affects all who are connected to the grid. As the system grows, reliance on it also grows, notwithstanding adversities such as environmental disruptions from burning non-renewable fossil fuels.

In contrast, methods of power generation and distribution have evolved as technology improved. Renewable energy generation through, for example, photovoltaic cells (solar power), not only encourages, but inherently drives a decentralized approach to power generation. Solar power generation is comparable to Fig. 1 b, where individuals or businesses generate their own energy, and only rely on utility power when demands increase above the attainable capacity.

There are many examples of technology driving decentralization, and in the modern era, technologies such as the internet-of-things, 3D printing, cryptocurrencies and decentralized urban farming are allowing citizens globally, in the developing and developed world, equal opportunities to contribute to economic growth. The following section reviews and discusses decentralized approaches in education and how they can benefit emerging markets and rural areas by excluding them from Industry 4.0.

IV. Re-evaluating Education in Emerging Markets

Over the last decade, many countries have emphasized widening access to higher education, which has led to rapid growth in the number of universities and degree programs on offer [11]. Sub-Saharan Africa, South Asia, and the Middle East and North Africa are among the nations with the lowest gross enrolment in tertiary education worldwide [12]. Decentralized approaches of tertiary institutions in both developed and developing countries have significantly scaled the uptake in enrolment in the last decade. Various prospects and challenges of an educational revolution that arises from the appropriation of digital technology into learning and educational practice are discussed in [13]. The following paragraphs identify additional strategies that re-evaluate how pedagogy is offered and the effects of implementing these strategies.

A. Satellite campuses

A growing sign of increased confidence and resourcefulness of universities in emerging markets is the growing number of branch (satellite) campuses being established. This is essentially a form of decentralization in

education and allows poorer learners unable to afford travelling to main campuses to enroll at universities.

B. Internationalization

Globally, higher education institutions are also developing new strategies to attract foreign students through teaching links with international partners to increase institutional visibility. Internationalization of higher education in the host country is not only effective in achieving a number of national aims [11], but also allows these countries to develop a workforce that meets the specific needs of their industries.

C. Standardizing the medium of instruction

The Academic Cooperation Association reports that since 2001 there has been an almost tenfold rise in the number of degree programs taught in English in countries where English is not the native language, the majority being at post-graduate level. As a result, the visibility, enticement, and approachability of these institutions have been raised for inbound students. Although at first glance this could resemble a centralized approached, the decentralization of this strategy is evident in the end result. Students gain knowledge in English and are capable of applying their knowledge and skills practically in their home country and native tongue. Furthermore, translation technologies have become supportive of student exchange and education (real-time narration).

D. Teaching through ICT

Collaboration and extending learning beyond the classroom is attainable through technology. Learning communities comprising learners, educators and external public and private services (such as libraries or even foreign universities) can be created and knowledge and information can be shared in a decentralized manner (as opposed to confining learning to a single prescribed textbook). In the digital classroom, the instructional approaches and techniques, the tools required, and adaptive skills and expertise of educators should complement such a collaborative strategy. Various technological aids can be applied in a modern digital classroom, and [1] identifies alternatives that are especially relevant and potentially beneficial in developing countries.

It is suggested in [1] that virtual and augmented reality-assisted teaching, learning, and training should be embraced and the huge potential of these offerings should be recognized and seized by educational institutions. Concepts such as numerical simulations or mechanical design can be simplified through developing a sense of interaction with the real world, and facilitating exploration and interpretation by learners.

Unlimited, open-access, non-discriminatory participation in courses can be facilitated through ICT, in developing more massive open online courses (MOOC), also supported by [1] and [11]. Interactivity and immediate feedback are among the enticing factors of these MOOCs, but more importantly for emerging markets, distance learning allows students from anywhere to access these course, provided ICT infrastructure exist. A level of digital literacy is required to access MOOC and language barriers could prevent learners from participation, but through regulated policies such as language standardization, MOOC could assist decentralization in education. In developing countries, governmental backing or funding from private investors, both domestic and international, could support this strategy. However, the cost of data in emerging economies is a barrier. A strategic decision towards zero-rating of websites or incentivizing businesses through a virtual economic zone could be a way forward.

E. Combining education with life-skills development

In [15], a new education model is identified that combines economic and social wellbeing of learners in developing countries with pedagogy. The work, based on multiple reviews of training programs in Asia, Latin America and Africa, is an approach that focuses on self-efficacy as a basis of employment and personal wellbeing, along with dynamic teaching in adaptive educational institutions worldwide [15]. It proposes uniting conceptual knowledge and practical experience that empower students to apply what they have learned. Despite being a well-known strategy in the developing world, this model is rarely applied in emerging markets (still focusing on content mastery), although a form of decentralized learning through internationalization is achievable through these methods.

V. Conclusion

In order to re-evaluate and scale the methodology of teaching learners in emerging markets the importance and potential of participating in Industry 4.0, an inherently different approach to education is required. Traditional centralized approaches are effective in certain circumstances, but inevitably evolving, and globalization and internationalization are crucial enablers to provide quality education for all. In this work, the role of ICT in education, as a driver for decentralization, is proposed and reviewed. The challenges and limitations in the developing world are vastly different when compared to developed countries, and often overlooked. The importance of a decentralized education system to scale programs, innovations, skills, and policies is argued in this work, recognizing the advantages and challenges, and providing motivations for governmental and private support in identifying the path towards socioeconomic sustainability.

VI. Research Recommendations

Sustainable and socioeconomic benefits from a scaled and decentralized education system are typically observed over long periods. Research into identifying the challenges and limitations of emerging markets that have a detrimental effect on long-term sustainability [4] is only the first step to mitigating their consequences. Active and constant research with annual analysis of progress by (typically non-profit) organizations is crucial in defining the feasibility of re-evaluating education in emerging markets.

References

[1] B. Zing, T. Marwala, "Implications of the fourth industrial age for higher education," in The Thinker, vol. 73, pp. 10-15, March 2017.

[2] S. O. Gunga, I. W. Ricketts, "The prospects for E-learning revolution in education: A philosophical analysis," in Educational Philosophy and Theory, vol. 40, no. 2, pp. 294-314, 19 May 2008.

[3] N. McGinn, T. Welsh, "Decentralization of education: Why, when, what and how?" in Fundamentals of Educational Planning, vol. 64, pp. 1-100, 1999.

[4] N. C. Nguyen, D. Graham, H. Ross, K. Maani, O. Bosch, "Educating systems thinking for sustainability: Experience with a developing country," in Systems Research and Behavioral Science, vol. 29, no. 1, pp. 1-16, 5 July 2011.

[5] K. Florestal, R. Cooper, "Decentralization of education: Legal issues," The World Bank, 30 June 1997.

[6] K. Sarabhai, P. Vyas, "The leapfrogging opportunity: The role of education in sustainable development and climate change mitigation," in European Journal of Education research, Development and Policy, vol. 52, no. 4, pp. 427-436, 24 October 2017.

[7] D. R. Winkler, A. I. Gershberg, "Education decentralization in Africa: A review of recent policy and practice," Draft (used by permission of the author), August 2003.

[8] A. Damon, P. Glewwe, S. Wisniewski, B. Sun, "What education policies and programmes affect learning and time in school in developing countries? A review of evaluations from 1990 to 2014," in Review of Education, Early View, 12 August 2018.

[9] I. A. Meyer, P. R. Gent, "The status of ICT education in South Africa and the way forward," Draft (used by permission of the author), 23 August 2016.

[10] D. Acemoglu, P. Aghion, C. Lelarge, J. Van Reenen, F. Zilibotti, "Technology, information, and the decentralization of the firm," in The Quarterly Journal of Economics, vol. 122, no. 4, pp. 1759-1799, 1 November 2007.

[11] University of Oxford, "International trends in higher education 2016-17," 2017. Available: http://ox.ac.uk. [Accessed: 4 Sept 2018].

[12] The World Bank, "Education Statistics (EdStats)," 2017. Available: http://datacatalog.worldbank.org. [Accessed: 4 Sept 2018].

[13] A. Collins, R. Halverson, "The second eduational revolution: Rethinking education in the age of technology," in Journal of Computer Assisted Learning, vol. 26, no. 1, pp. 18-27, 12 January 2010.

[14] Q. Zhou, H. Sun, R. Zhou, G. Sun, J. Shen, K. Li, "A collaborative and open solution for large-scale online learning," in Computer Application in Engineering Education, pp. 1-16, 7 August 2018.

[15] M. J. Epstein, K. Yuthas, "Redefining education in the developing world," 2012. Available: http://ssir.org. [Accessed: 1 Sept 2018].

Fabless Semiconductor Implementation Model as an Enabling Factor for "Virtual Labs" in Online Postgraduate Degrees in Microelectronics

Mladen Božanić, *Senior Member, IEEE*
Faculty of Engineering and the Built Environment
University of Johannesburg
Johannesburg, South Africa
mbozanic@ieee.org

Saurabh Sinha, *Senior Member, IEEE*
Deputy Vice-Chancellor, Research and Internationalisation
University of Johannesburg
Johannesburg, South Africa
dvc-r@uj.ac.za

***Abstract*—Modern technology and the increasingly readily available access to the internet in recent years have initiated the paradigm shift from the "local student" to the "global and virtual student", which is evident in the increased numbers of students who are enrolling in the postgraduate (master's and PhD) degrees offered through remote means. Such students are not necessarily even present in the town or the country of the university at which they have enrolled. Traditional walk-in universities such as the University of Johannesburg, South Africa, are consequently increasing the portfolio of online degrees, which is in line with the trend seen by many universities worldwide. Education in electronics traditionally requires a hands-on approach, and the laboratory component remains a requirement even at postgraduate level. In this paper, the fabless semiconductor manufacturing approach, already deployed by many microelectronic industry giants, is aligned as an enabling factor of establishing "virtual labs", used to fulfil the practical education component of both research-based and coursework microelectronic modules offered online. A case study involving postgraduate degrees with specialization in microelectronics from the University of Johannesburg is also presented.**

Keywords—Postgraduate education, master's degree, doctoral degree, microelectronics, virtual labs, fabless semiconductor manufacturing

I. INTRODUCTION

In recent years there has been an observable increase in student enrolment figures, at least when postgraduate education is considered. Given socio-economic considerations, students are increasingly attracted by degrees that are offered part-time [1]. Many of such students are interested in distance education. South Africa is known to have one of the largest distance education tertiary institutions – the University of South Africa (UNISA) [2]. In addition, the increasingly readily available access to the internet and the decreasing cost of devices with this connectivity in fact allowed for the emergence of the online model of distance education, which has resulted in increased interest by many universities, whether traditional walk-in ones or distance-based. In the "walk-in" approach, the quality of education delivery is being supported through a blended approach; on the other hand, the online approach is contributing to increased accessibility.

Online enrolment means that that both lecturing and assessment are done using specialized education platforms (for example Blackboard Learn [3]) and students are not required to be present in person at the university, not even for examinations (which allowed the term "virtual student" to be coined). This differs from typical distance education where students are instructed using the specialized platforms only, but are required to travel for examinations and practical work (as in the case of many degrees offered by UNISA). The online approach results in minimum costs of study, with no lodging and commuting cost for students and no costs for lecture halls, material and equipment for the university.

Examination is still, naturally, required for all degrees, and it is furthermore difficult to avoid the practical component in degrees that require a hands-on approach, as in the case of traditional degrees in electrical and electronic engineering. Fortunately, purely online degrees are a better fit for postgraduate (graduate) education, where traditional (written or oral) examination can be replaced by examination assignments or continuous assessment, which requires unique solutions from each student. This is possible because the focus of education at postgraduate level has shifted from analytical to research-based or specialist education. Aiding this is the fact that student enrollment numbers are much lower than in the case of undergraduate degrees. This leaves only the requirement for a typical degree to have a practical component as the last step in allowing for the concept of a truly "virtual student" to emerge, at least in the field of electrical engineering, and at least at graduate level. The answer can possibly be sought in the field of microelectronics and aligned to this approach, which is already led by the same industry.

Microelectronics, and more recently nanoelectronics, are two related fields of electronic engineering typically taught at postgraduate level. These can both be defined as microtechnology applied in the context of electronic circuits and systems [4], where manufacturing in the sub-one-μm (sub-one-micron) process is termed nanoelectronics by some researchers and industry professionals. Microelectronic design in industry is typically performed using specialized design and simulation software packages, called electronic design automation (EDA) packages [5], which typically run on powerful computer servers, accessed through computer terminals. This is different from electronic design in other disciplines, where the actual components and laboratory equipment are needed in all design steps. The end product of EDA-aided designs is a set of blueprints for semiconductor integrated circuit (IC) fabrication [6]. This makes microelectronics one of the disciplines of electronic engineering where an almost complete design cycle can be completed from the comfort of an office, or even home office, with the workstation practically serving as a virtual laboratory. The remaining steps in the design and manufacturing cycle are the fabrication of semiconductor

wafers based on created blueprints, as well as testing, packaging and characterization of individual ICs that appear on manufactured wafers, both of which are done in specialized, expensive industry facilities. The fact that manufacturing (fabrication) and test facilities are extremely expensive has driven the emergence of fabless semiconductor design companies that rent capacity from third-party fabrication facilities and test houses instead of building their own, to complete the implementation cycle.

In this paper it will be shown how the fabless manufacturing industry concept can be used in research and education. Universities can, like fabless semiconductor manufacturing companies, utilize the same process where students perform research and design in the same specialized EDA software industry individuals would use. Since the virtual laboratory consisting of the set of EDA tools can be accessed remotely, this form of education is in line with the emerging online education trend.

The paper is organized as follows: Fabless semiconductor manufacturing is described from the industry perspective in Section II. In Section III, various types of microelectronic postgraduate degrees presented globally are researched. In Section IV, a case study of the University of Johannesburg (UJ) is used to show how different online modes of delivery can benefit from the practical component delivered by the fabless semiconductor model. Finally, Section V concludes the paper.

II. Fabless Semiconductor Implementation: The Industry Model

To understand the importance of the IC and the flow of semiconductor implementation, it is necessary to touch briefly on the history of the IC.

Soon after the transistor was invented in 1947, it became clear that if semiconductors were to reach commercialization, they would need to be greatly miniaturized [7, 8]. It was only in 1959, however, when it became possible to embed electrical connections onto silicon, that the idea of an IC was conceived. By the mid-1980s, all silicon ICs were produced by companies that had their own fabrication facilities (fabs) – the companies that are now called integrated device manufacturers (IDMs) [9, 10]. This advanced the Third Industrial Revolution. Fabs were not cheap, but they could be afforded by relatively small companies, for several reasons. For one, manufacturing processes by the laboratories were relatively cheap, because of the then large chip features (dimensions). Secondly, required wafer volumes were small, so small fabs with low maximum volumes were sustainable. This started changing in the 1980s, when processes reached feature sizes that made processing expensive, and consequently the volumes required to keep the fabs running increased drastically. It then became evident that the sharing of fabrication costs was the answer to sustain semiconductor manufacturing. In early efforts at sharing fabrication costs, certain companies would rent spare fabrication capacity from an IDM. In 1987, finally, Taiwan Semiconductor Manufacturing Company (TSMC), the first semiconductor company that did not produce its own ICs at all, but rather rented its fabrication space, was established. Using design models and kits from TSMC, any corporation could now design ICs that would be fabricated at the foundry for a fee, and fabless semiconductor manufacturing, as it is known today, was born. Today, a large number of fabrication facilities exist, and many semiconductor industry leaders, such as Apple, Qualcomm and Nvidia, use the fabless model of manufacturing.

Typical modern fabless implementation flow follows the steps in Fig. 1. Designs are prepared using specialized software and verified in simulation, after which layouts (blueprints) for fabrication are created. Blueprints are used to create masks for different fabrication layers. In the fabless model, the fabrication step in Fig. 1 is outsourced, typically to the Far East where costs can be reduced further. Upon fabrication, circuits are tested and packaged (last step).

The most critical implementation step for fabless corporations is the design phase. To ensure first-pass success, various design and verification steps need to be put in motion by the design team, as illustrated in Fig. 2 (adapted from [11]). Two separate phases can be identified: system design, where top-level circuits are completed in design software, using libraries and process development kits (PDKs) supplied by the foundry; and in the second phase, blueprint generation, the design is converted into a format that can be interpreted by the semiconductor foundry, using the foundry process design rules. Both phases rely heavily on the use of EDA, as stipulated in the introduction of this paper. Depending on the complexity of the integrated system and types of simulations that need to be performed, various combinations of tools listed in Table I (or any others not specified here) can be used by the design teams [12].

While the fabless approach has a prototyping advantage, the designer or mindset has the risk of detaching from the aspect of "designing for testability (DFT)." To incorporate DFT, the approach must include, among others, process, voltage and temperature variation simulations and also package-based simulations. In the last-mentioned, pads are simulated as a circuit or system component.

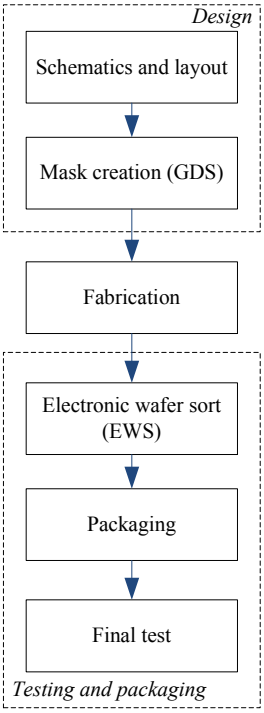


Fig. 1. Microelectronic industry implementation flow.

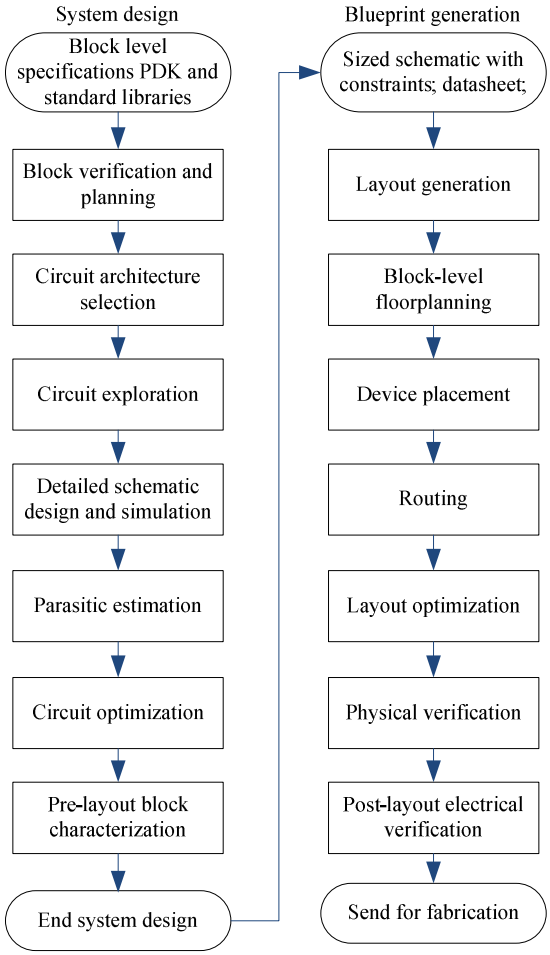


Fig. 2. Most important steps in the typical microelectronic system design flow using EDA, from block planning to system implementation.

TABLE I. TOOLS COMMONLY USED IN MICROELECTRONIC DESIGN

Tool name	Vendor	Type of simulation possible
Microwave Office	NI AWR Corporation	Linear and nonlinear EM simulation
Advanced Design System (ADS) with Momentum and EMPro	Agilent Technologies Inc	Full linear and nonlinear system design (including layouts)
IE3D	Zeland Software Inc. (acquired by Mentor Graphics)	3-D EM simulations
CST Microwave Studio	Computer Simulation Technology AG	3-D EM simulations
Ansoft HFSS	ANSYS Inc.	3-D EM simulations
Virtuoso with Analog Design Environment (ADE) and SpectreRF	Cadence Design Systems, Inc.	Analog, digital and RF IC design with non-linear simulations (including layouts)
Sonnet	Sonnet Software	IC EM modeling and simulation up to THz frequencies
SmartSpiceRF	Silvaco	Physical-level device simulation and netlist-based design and layout
Freeware tools	Various vendors	Rudimentary level design, simulation and layouts

III. POSTGRADUATE EDUCATION IN MICROELECTRONICS

To relate the fabless semiconductor manufacturing to education in the field of microelectronics, the approach taken by top-ranked tertiary institutions to education in microelectronics and related fields will briefly be reviewed. Doctoral degrees are still most often offered by research; master's degrees, however, can be offered as three types, listed in Table II [13].

A. Master's by Research

Master by research is what is generally considered as the traditional master's degree. It is similar to the doctoral degree in that an independent study must be completed by the student, but the level of contribution to the body of knowledge is minor compared to a doctorate. Many top-ranked institutions still offer research master's degrees, including Cambridge, Oxford and Harvard. In certain instances, students can complete the complete degree without being physically present at the university, provided that the practical component is honored sufficiently.

TABLE II. MASTER'S DEGREES WITH A MICROELECTRONIC COMPONENT AS OFFERED BY TOP-RANKED INSTITUTIONS

Degree type	University	QS - Ranking (2016)	Degree name
Research	University of Cambridge (Cambridge, UK)	4	MPhil in Engineering
	Harvard University (Cambridge, MA, US)	9	Master of Engineering
	University of Oxford (Oxford, UK)	11	MSc by Research in Engineering Science
Coursework	Stanford University (Stanford, CA, US)	2	MS Electrical Engineering
	University of California (Berkeley, CA, US)	3	Advanced Study in ICs (online)
	University of California (UCLA) (Los Angeles, CA, US)	5	MS in Engineering (online)
	National University Singapore (Singapore)	6	MSc Electrical Engineering
	Nanyang Technological University (Singapore)	8	MSc Microelectronics
	Georgia Institute of Technology (Georgia Tech) (Atlanta, GA, US)	14	MS Electrical and Computer Engineering
	California Institute of Technology (Caltech) (Pasadena, CA, US)	12	MS Electrical Engineering
Combined	Massachusetts Institute of Technology (MIT) (Cambridge, MA, US)	1	Master of Engineering
	ETH Zürich (Zürich, Switzerland)	8	Master's Electrical Eng. and IT
	Imperial College London (London, UK)	10	MSc Analogue and Digital IC Design
	Swiss Federal Institute of Technology in Lausanne (Écublens, Switzerland)	13	MS Electrical Engineering (Electronics and Microelectronics)

B. Master's by Coursework

In this mode of postgraduate education, students enroll in a number of compulsory and elective modules, which are credited to a complete specialization. Coursework degrees are often deployed in fields of study where the workload at undergraduate level is too broad for specialized modules to be offered. Microelectronics, as a specialization of electronic engineering, can often be taught at master's level (some universities in the United Kingdom and South Africa, among other countries, offer honors degrees in microelectronics, allowing for specialization). Unless the module is designed to be offered online, students' presence on university campuses is mandatory. Stanford, Berkeley, UCLA, the National University Singapore, Nanyang Technological University, Georgia Tech and Caltech are top-ranked universities that offer coursework master's degrees.

C. Master's by Coursework with a Mini-Dissertation

In recent years, some institutions commenced splitting the total credits towards degree completion between the coursework and research component. This way, the benefits of both coursework and research master's degrees can be exploited. Normally, core modules are used for student specialization, after which students are required to complete a research proposal as well as a small research dissertation. This model is followed by MIT and some European top-ranked institutions.

D. Doctoral Degrees

Finally, doctoral degrees are typically completed in research mode. Students enrolling in research doctoral degrees are normally disciplined, which allows them to complete their theses with even less presence on university campuses.

IV. Case Study: Fabless Approach to Practical Component in Postgraduate Degrees with Specialization in Microeletronics at the University of Johannesburg

Microelectronics can be studied at both master's and doctoral levels at UJ. Traditional master's and doctoral qualifications offered by the Faculty of Engineering and the Built Environment (FEBE) are research-based "walk-in" qualifications, where microelectronic topics can be chosen with supervisors from the Department of Electrical and Electronic Engineering Science. Although research degrees are not formulated as online degrees, UJ has a number of contracts with EDA vendors that allow students to complete most of their work using the online model. There is also an agreement with MOSIS Integrated Circuit Fabrication Service and UJ has benefited from a prototyping sponsorship in 2018. In this case, MOSIS coordinates with GlobalFoundries.

The extent of the work differs between the master's and doctoral qualifications. The full potential of the fabless approach, however, will be exploited at UJ from 2019, when a new, completely online master's degree, approved in 2018, will be offered under the name of Master's in Micro and Nano-Electronic Engineering, the curriculum of which is developed in line with modern trends in microelectronic education and industry, as will be described shortly.

The extent to which the fabless semiconductor manufacturing model can be exploited in each of the abovementioned degrees will be explored in the sections that follow.

A. Master's by Dissertation

Typically, research conducted in traditional postgraduate degrees is already conducted by the student in isolation (whether on campus or off campus), with some support by the supervisor through a limited number of contact sessions. At UJ FEBE, master's degree supervision is primarily conducted using online communication tools, and general support to students will be offered via the existing university structures, such as the Centre for Academic Technologies (CAT), as well as library and administration staff. CAT supports the Blackboard learning management system and, as of 2017, advanced Blackboard Predict analytics has also been available. The approach of predictive analytics is part of the university's quest for the Fourth Industrial Revolution (Industry 4.0) and embedding the approach also through pedagogical practices.

For the electronic engineering master's degree by research study, research into a microelectronic topic of students' choice can be conducted using EDA available to UJ, which can be accessed remotely. For the master's dissertation, students are typically required to develop a design prototype to verify the research questions posed in the research proposal. Simulation capabilities of the EDA tools allow for the prototype to be designed and simulated. Students are typically also required to complete layouts, which are free of errors and ready for manufacturing in an outsourced fab. Designs are completed using real-life PDKs available to the university, provided via an agreement between UJ and MOSIS Integrated Circuit Fabrication Service (coordinating with commercial foundries). On the master's level, fabrication is required for walk-in students; student's presence on one of the campuses is required only to complete prototype measurements. Given the developing level of specialization, some measurement work is also being done through international partnerships. For a tentative online research master's degree, for students based outside the greater Johannesburg area, a possibility to perform measurements at partner institutions close to their place of residence could be investigated.

Authors of this paper are currently supervising/co-supervising at least one master's student using the fabless approach, as described above.

B. Master's in Micro- and Nano-Electronic Engineering (Online)

The duration of the study for the new online master's degree in Micro- and Nano-Electronic Engineering will be one year (two semesters). The requirements for degree completion will be the two core modules, two electives, the Engineering Research Proposal Writing module, and a mini-dissertation. The electives are chosen from a pool of electives, and the pool of electives prepared for pilot deployment which will commence in 2019 includes Power Electronics, Optoelectronics and Advanced Semiconductor Manufacturing, whereas cross-disciplinary specialization can be obtained by enrolling in online modules in other faculties or even institutions. Additional elective modules are also still under development as of 2018 [14]. It is also possible to

develop a model of a joint qualification where certain modules are completed at partner institutions [15].

As is the case with the research master's degree, degree supervision will primarily be conducted using online communication tools, and general support to students will be offered via the existing university structures. The mini-dissertation in the degree will be dealt with in a similar manner as the dissertation of the master's by research, with the exception that open-access EDA tools will be favored. UJ has a general preference and policy favoring open-access academic material. Also, even though camera-ready layouts could be expected from the student based on the topic of research, because of the long time required for the fabrication cycle to be completed (typically six months), actual fabrication will not be justified. The online model will allow both full-time and part-time university personnel to be involved in supervision, thus opening up the opportunity for industry or science council experts to be brought in to assist. Core and elective modules will also have a strong fabless research component, comprising typically schematic and hardware-level design as well as simulation. A more detailed description of the extent of the fabless approach in each of the discussed core modules and electives can be found in Table III.

C. Doctoral Degree with Specialization in Microelectronics

Lastly, the doctoral degrees follow the model of master's by research, with the complexity of research typically requiring many different EDA tools to be used in conjunction, with the complete fabless flow deployed as in Fig. 1 and Fig. 2. As leading engineering journals normally require measurement results for article publication (and publication of articles is one of the requirements for doctoral degree completion), fabricated prototypes have to be measured. This can, however, typically be facilitated by available doctoral research grants. Such approach has been utilized by a recent PhD graduate supervised/co-supervised by the authors, thus proving the effectiveness of the said concept.

TABLE III. DEGREE COMPONENTS, MODULES AND THE EXTENT OF THE FABLESS APPROACH DEPLOYED IN EACH OF THE MODULES OF THE MASTER'S IN MICRO- AND NANO- ELECTRONICS OFFERED BY UJ FEBE

Degree component	Module name	Extent of the fabless design approach
Core modules	Analog and RF Microelectronics	Analog and RF schematics, analog and RF simulations
	Digital Design and Synthesis	Hardware-description language design, simulation and synthesis
Elective modules (as of 2019)	Power Electronics	Power electronics schematics, power simulation
	Optoelectronics	Optoelectronic design and simulations
	Advanced Semiconductor Manufacturing	Layout tools; layout analysis
	Cross-disciplinary elective	-
Research component	Research Proposal Writing	-
	Mini Dissertation	Analog, digital and RF design, simulation, synthesis, place and route, layout

V. CONCLUSION

In this paper, it was discussed how parts of the fabless semiconductor implementation flow can be used in online microelectronic education to fulfill the requirement for the practical component. This approach is already used to a great extent in many research degrees with a microelectronic component, but the requirement for the prototype measurement usually means that students need to use the actual laboratories towards degree completion. UJ's Master's in Micro- and Nano-Electronic Engineering degree that will start in 2019, which combines coursework with a mini-dissertation component, however, was developed in such a way that the whole curriculum can be covered by using only the "virtual" portion of the fabless semiconductor manufacturing flow, thus introducing "virtual labs" into the degree. Future research will involve the analysis of the degree of success of the fabless education approach in this specialized degree, but this will only be possible after the degree has been run for at least two years.

ACKNOWLEDGMENT

MOSIS Integrated Circuit Fabrication Service and EDA vendors, such as Keysight, have contributed to postgraduate and research work. MOSIS, through the MOSIS Educational Program, has supported coordination with leading foundries and thus given students the opportunity for industry-relevant research. Keysight has contributed through heavily discounted simulation software licensing and in particular usage of ADS.

REFERENCES

[1] C. Murarane za, F. Mtshali, F., and S. Z. Mthembu, "Research supervision: Perceptions of postgraduate nursing students at a higher education institution in KwaZulu-Natal, South Africa," *Afr. J. Health Professions Educ.*, vol. 8, no. 2, pp. 135-139, Oct. 2016.

[2] T. V. Manyike, "Postgraduate supervision at an open distance e-learning institution in South Africa," *South Afr. J. Educ.*, vol. 37, no. 2, pp. 1-11, May 2017.

[3] W. Rice, *Blackboard Essentials for Teachers*. Birmingham: Packt Publishing, 2012.

[4] E. Bruun, D. Demarchi, and I. R. Nielsen, "European master programs in nanoelectronics and microsystems," in *Proc. of the 10th Eur. Workshop Microelectron. Educ., 14-16 May 2014, Tallinn, Estonia* [Online]. Available: IEEE Xplore, https://ieeexplore.ieee.org/abstract/document/6877394. [Accessed: 1 Oct. 2018]. pp. 53–57.

[5] R. Brayton, L.P. Carloni, A.L. Sangiovanni-Vincentelli, and T. Villa, "Design Automation of Electronic Systems: Past Accomplishments and Challenges Ahead," *Proc. IEEE*, vol. 103, no. 11, pp. 1952-1957, Nov. 2015.

[6] A. Hastings, *The Art of Analog Layout*, 2nd ed. Upper Saddle River: Prentice Hall, 2006.

[7] V. Smil V., "July 1958: Kilby conceives the IC," *IEEE Spectrum*, vol. 55, no. 7, p. 22, July 2018.

[8] P.M. Raj, H. Sharma, S. Sitaraman, D. Mishra, and R. Tummala, "System Scaling With Nanostructured Power and RF Components," *Proc. IEEE*, vol. 105, no. 12, pp. 2330-2346, Dec. 2017.

[9] R. Kumar, *Fabless Semiconductor Implementation*. New York, NY: McGraw-Hill, 2008.

[10] D. Nenni, and P. McLellan, P., *Fabless: The transformation of the semiconductor industry*, SemiWiki LLC, 2013.

[11] J. Chen, M. Henrie, M. F. Mar, and M. Nizic, *Mixed-Signal Methodology Guide*, San Jose: Cadence Design Systems, 2012.

[12] M. Božanić, and S. Sinha, S., *Millimeter-Wave Low Noise Amplifiers*, Cham: Springer Nature, 2018.

[13] M. Božanić, and S. Sinha, "A survey of current trends in master's programs in microelectronics," *IEEE Tran. Educ.*, vol. 61, no. 2, pp. 151-157, May 2018.

[14] M. Božanić, S. Chaturvedi, and S. Sinha, "Re-inventing postgraduate level teaching and learning in nanoelectronics," in *Proc. of the 2017 IEEE AFRICON, 18-20 Sept. 2017, Cape Town, South Africa* [Online]. Available: IEEE Xplore, https://ieeexplore.ieee.org/abstract/document/8095564. [Accessed: 1 Oct. 2018]. pp. 676-681.

[15] M. Božanić, S. Chaturvedi, and S. Sinha, "Evolving the BRICS multi-lateral qualification through online education innovation," in *Proc. of the Int. Conf. Empowerment through Educ., Ethics & Entrepreneurship (IC-4E) - The Next Agenda for BRICS*, *3-4 Feb. 2017, Mandsaur, India.* Mandsaur: Mandsaur University. pp. 25-29.

Comparison Between the e-Learning Readiness of Educators and Learners in South African Schools

Irene Kolo and Tranos Zuva
Information and Communication Technology
Vaal University of Technology
Vanderbijlpark, South Africa
irene.kolo@gmail.com, tranosz@vut.ac.za

***Abstract*—In this paper, e-learning readiness of the South African schools in previously disadvantaged schools in Gauteng were investigated. A twenty-nine (29) item questionnaire was used to obtain the data. The new learning method is being introduced and will be implemented by Gauteng Department of Education. The data obtained from the questionnaires was then analyzed by using a STOPE model. The results obtained by using STOPE show that the learners are 78% ready as compared to 74% of educators. The five-point likert scale showed a 3.91 average learners' readiness as well as a 3.69 average educator's readiness. Thus, it is evident that the educators are generally less receptive to adopting e-learning than the learners.**

Keywords—Comparison, e-learning readiness, educators, learners

I. INTRODUCTION

In the present day, technology is used for collaboration amongst educators and learners. The South African government is currently pushing to roll out e-learning in South African schools[1]. E-learning is used to deliver education curriculum through the latest available technologies. Some schools in South Africa are using a combination of traditional and e-learning methods, while some are still using the traditional way of delivering the curriculum. To build an effective e-learning environment many things need to be prepared, such as the readiness of technology, the readiness of educational institutions, and the readiness of the community. These preparations are important because they will affect the quality of e-learning programs when implemented [2].

E-learning has gained popularity in higher education because learning technologies help to address many of the key challenges that colleges and universities face [3]. There are many benefits offered by e-learning which include advantages of convenience, time and flexibility, no need to be at the campus and more advanced information[4, 5]

One of the realities is that, while all students can learn, they will not all start at the same place, learn at the same rate, or reach the same ultimate level of proficiency. This is because of differences in academic ability, interest, and commitments [6]. This study will determine in which areas educators and learners are not yet ready for e-learning, for example, technology skills and attitude towards e-learning.

The primary objective of this study was to assess the readiness for e-learning in participating South African schools in Gauteng province to assure successful e-learning implementation. The secondary objectives of this research were to study previous papers on e-learning readiness, to propose an appropriate e-learning readiness model for assessing the readiness of e-learning and, to measure e-learning readiness in South African schools.

This paper proceeds as follows: In section II, we review some of the existing models that are used to assess e-learning readiness. The methodology is provided in section III, while the results are discussed in section IV. Finally, we conclude the paper in section V with a brief note on future work.

II. RELATED WORK

This paper reviews literature on e-readiness from previously written papers from different countries, defines e-learning and summarizes the different readiness models.

It was confirmed by [7] that there has been disputes about a common definition of the e-learning. According to [7], e-learning is "the use of information and communication technologies to enable the access to online learning/teaching resources." The models and frameworks used in other countries reviewed and assessed. Models and frameworks developed with similar characteristics to South African schools in mind, were then used to assess the e-learning readiness in South African schools. The STOPE model is discussed in detail in methodology.

Table I summarizes the previously developed models used to measure the readiness of e-learning in different countries. "Most of the models were developed for use in business organizations, universities or higher education institutions. In addition, they were designed for use in developed countries whose e-maturity is high. Every system, (organization, culture, country and individual) has its own norms, for that measurement instruments that work in one country might not work for organizations in other countries" [8].

TABLE I. DIFFERENT PREVIOUSLY USED MODELS

Author	Factors
[9]	1. Business 2. Technology 3. Content 4. Culture 5. Human resources 6. Financial resources
[10]	1. Content 2. Pedagogy 3. Technology 4. Culture 5. Awareness
[11]	1. Strategy 2. Structure

	3. Systems 4. Style/culture 5. Staff 6. Skills 7. Shared values

III. METHODOLOGY

This study used quantitative. Data was collected from educators and learners using closed-ended questionnaires. The questions on the questionnaires were adopted from previous papers written by [12, 13]. Descriptive method was used to determine the level of readiness to implement e-learning in South African schools. For the level of readiness measurement, a five-point Likert scale was used.

For the purpose of this study, the structured questionnaires were distributed to the educators and learners in different high schools in Gauteng province. Each individual educator or learner was expected to fill in their own questionnaire individually. For the learners, consent forms were distributed and only those who had signed forms were administered the questionnaires for participation. Participants and non-participants were not harmed. The approval was received from Gauteng Department of Education for the schools to participate. The province consists of 667 public schools and 1 325 969 learners[14].

Targeted participants were both females and males between the age of 13 to 50. A population of 176 participants was reached. Out of 176, only 48 was educators and 128 learners from grade 7 to grade 12. The main research question of this study was: how can assessment of e-learning be done in South African schools to assure success in implementing e-learning?

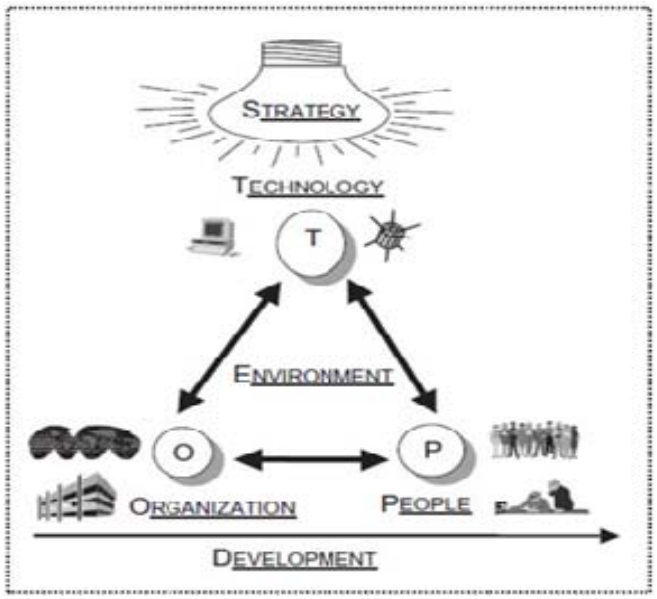


Figure 1. STOPE approach [15]

The elements of STOPE are identified as strategy(ICT directions and plans), technology(ICT facilities), organization(ICT rules and management), people(ICT skills) and environment(non-ICT issues) [16].

In the proposed model the STOPE will be:

Strategy = Strategy, Technology = Technology, Organization = Institution, People = People and Environment = Content

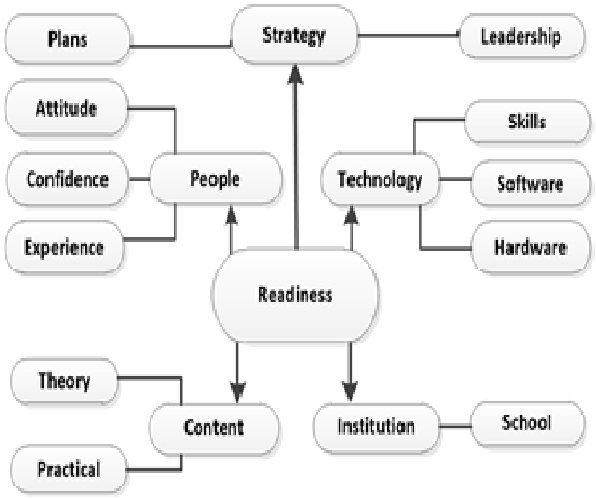


Figure 2. The proposed model

The proposed model above (figure 2) was intended for both learners and educators and it can be applied for different age groups. The consent forms will still be essential for the participants under the age of 18 due to ethical considerations. The sampled leaners used in this study were all from high schools and their ages ranged from 13 years to 20 years. Educators' ages ranged from 28 to 50 years.

The issues were addressed by adapting the previous models developed for measuring readiness. The adapted model was validated by educators and leaners participation in the survey administered.

Table II below illustrates the five-point Likert scale meanings.

TABLE II. FIVE-POINT LIKERT SCALE MEANINGS

Means	Scale
1 - 2.6	not ready lot of work to be done
2.6 - 3.4	not ready some work to be done
3.4	expected level of readiness
3.4 - 4.2	ready but needs a few improvements
4.2 - 5	ready to go

Table III below explains purpose of the components included in the proposed model. The table is consisting of five domains namely: strategy, technology, institution, content and people. The domains contain 25 factors distributed over 11 issues.

TABLE III. DISCUSSION OF PURPOSE OF EACH VARIABLE IN THE MODEL

Readiness			
Variable	**Issues**	**Aim**	**Factors**
People	Attitude	To determine the attitude	2
	Confidence	To understand the feeling	3

		that one can rely on e-learning	
	Experience	To determine the knowledge acquired	5
Technology	Skills	To determine the technological skills	4
	Software	To check the available software	2
	Hardware	To check the available equipment	3
Content	Practical	To determine the usage of e-learning	1
	Theory	To determine general principles	1
Strategy	Leadership	To assess the action of leading an organization	1
	Plan	To assess the intention	2
Institution	School	To determine if e-learning is used	1

IV. RESULTS

This paper discusses the findings and analyses from the consent forms signed by the parents of the learners studying in certain section of Gauteng high schools. The purpose of this study was to assess e-learning readiness in South African schools to assure success in implementing e-learning. The objective of this study was to determine the e-readiness in South African schools to assure success and quality.

The questionnaire consisted of 5 sections. The sections are: strategy (3), technology (9), institution (1), people (10) and content (2).

Below are the results of each issue for the learners and educators separately. The results obtained compared as follow:

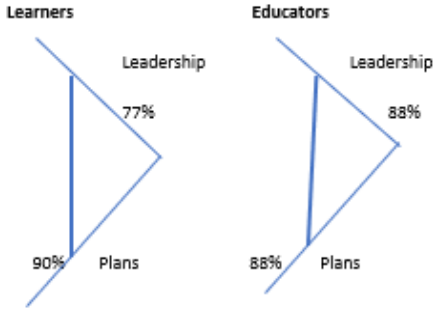


Figure 3. Strategy domain comparison assessment results

a. The Strategy Domain

Figure 3 above summarizes the results attained for the e-readiness assessment issues of the strategy domain. The issues used to assess the strategy domain were plans and leadership.

In terms of plans, participants were asked if they were willing to spend money because of e-learning and if they want to be involved in e-learning. The results indicated 90 percent and 88 percent of learners and educators respectively, are willing to spend more money on e-learning. On the issue of involvement, 88 percent and 91 percent of learners and educators respectively, want to be involved in e-learning.

In terms of plans as a measure on both learners and educators, 90 percent and 88 percent of learners and educators respectively agree with e-learning, see figure 3 above. In terms of leadership the participants were asked if computers are adequately provided for e-learning, 77 percent of the results indicated that learners believe that computers are provided, while 88 percent of educators agree.

The strategy domain as measured by both plans and leadership showed 84 percent of leaners agreeing while educators are at 88 percent.

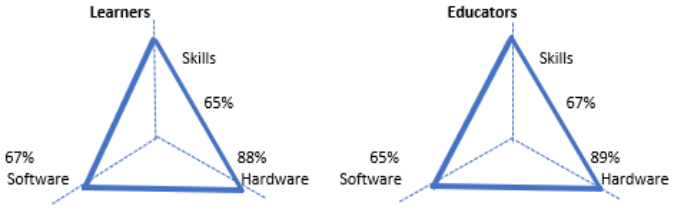


Figure 4. Technology domain comparison assessment results

b. The Technology Domain

Figure 4 above summarizes the results attained for the e-readiness assessment issues of the technology domain. The issues used to assess the technology domain were skills, hardware and software.

In terms of the skills the participants were asked if they know how to resolve basic hardware or software problems, how to download files using different browsers, how to resolve common errors while surfing the internet and how to navigate the web pages. The results indicated that 64 percent of learners believed they have basic computer skills while 67 percent of educators believed the have basic computer skills in navigating through different computer programs and to resolve certain software and hardware systems.

In terms of hardware the participants were asked if they can easily get access to a computer in the ICT centre, if tablets are adequately provided and if the IT infrastructure in their school can support e-learning. The results indicated that 88 percent and 89 percent of students and educators respectively basic hardware facilities such as computers and tablets were provided and they have access to such facilities.

The participants were also assessed in software knowledge and accessibility. The questions asked included access to web and software provision. The results indicated that 67 percent and 65 percent of students and educators respectively have access to web and software. The technology domain as measured by skills, hardware and software showed 73 percent for both leaners and educators.

In terms of software the participants were asked if they can easily get access to web and if software is adequately. 65 percent of the results indicated that software resource is readily available. The remaining 35 percent did not agree that software resource is readily available.

The assessment in terms of technology domain indicated that on average technology is 73 percent readily available to be used in the form of e-learning.

Figure 5. Institution domain comparison assessment results

C. The Institution Domain

Figure 5 above summarizes the results attained for the e-readiness assessment issues of the institution domain. The issue used to assess the institution domain was school readiness to accept the e-learning.

In terms of the institution domain, the participants were asked if their schools were willing to accept e-learning as a mode for teaching and learning.

The results indicated that 93 percent and 79 percent of learners and educators respectively, believed that their schools are willing to accept e-learning as a mode of offering education in their schools.

The institution as measured by theory and practical showed 93 percent and 79 percent for both leaners and educators respectively

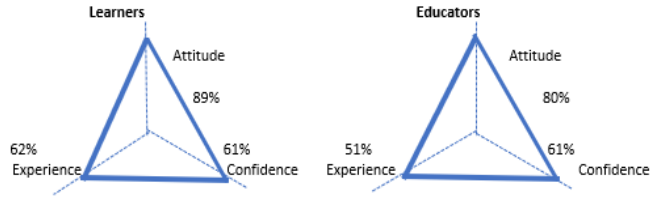


Figure 6: People domain comparison assessment results

D. The People Domain

Figure 6 shows the results attained for the e-readiness assessment issues of people domain. The issues used to assess the people domain were attitude, experience and confidence.

In terms of the attitude the participants were asked if they believe that using e-learning can increase their productivity and if they believe that e-learning will enable them to accomplish their learning and/or teaching more effectively than the traditional classroom-based approach. The results indicated that 89 percent and 80 percent of students and educators respectively have positive attitude towards e-learning.

In terms of experience the participants were asked if they owned a laptop/desktop, if they owned an email address and can open/ send and add attachment(s), if they were able to minimize and maximize opened application(s), if they were familiar with Microsoft Office and if they could differentiate between .pdf, .docx., .xls and .ppt. The results indicated that 61 percent of both leaners and educators have basic experience of to operate a computer.

In terms of confidence the participants were asked they believed that e-learning can improve the quality of learning and/or teaching, if they believed that it is easy for them to use e-learning, if they believed that leaners and/or educators will find it easy to use e-learning. The results indicated that 62 percent and 51 percent of learners and educators respectively, have confidence that they can use e-learning.

The people domain as measured by attitude, confidence and experience showed 71 percent and 64 percent for both leaners and educators respectively.

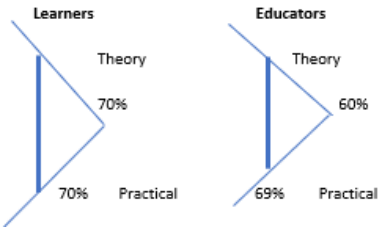


Figure 7: Content domain comparison assessment results

c. The Content Domain

Figure 7 above summarizes the results attained for the e-readiness assessment issues of the content domain. The issues used to assess the content domain were theory and practical.

The results indicated that learners believed that 70 percent of the e-books were adequately provided while educators showed 60 percent. In terms of practical learners believed that 70 percent of interactive resources like whiteboards were provided, educators believed that 69 percent of interactive resources such as whiteboards were provided.

The content domain as measured by theory and practical showed 70 percent and 64 percent for both leaners and educators respectively.

TABLE IV. STIPC = STOPE WEIGHTED INDICATOR

Learners			
Domain	M (1-5)	0.21	Ind. (%)
S	4.19	0.19	84%
T	3.66	0.18	73%
P	3.54	0.18	71%
C	3.50	0.24	70%
I	4.65	1.00	93%
Total	19.54	0.20	84%
Overall	3.91		
Indicator	78%		
Educators			
Domain	M (1-5)	W (1)	Ind. (%)
S	4.4	0.24	88%
T	3.67	0.20	73%
P	3.21	0.17	64%
C	3.22	0.17	64%
I	3.96	0.21	79%
Total	18.46	1.00	
Overall	3.69	0.20	
Indicator	74%		

Table IV above consists of five domains, namely; strategy (S), technology (T), people (P), content (C) and institution (I). The measure started from 1 to 5 as there are five domains to be measured. The indicators were obtained

from the respondents on the questionnaires, and they were used to calculate the measure of each domain, ranging from 1 to 5. The last step was to calculate the weighted average of each domain, using the measurements obtained. The readiness indicator is between 78% and 74% for learners and educators' respectively.

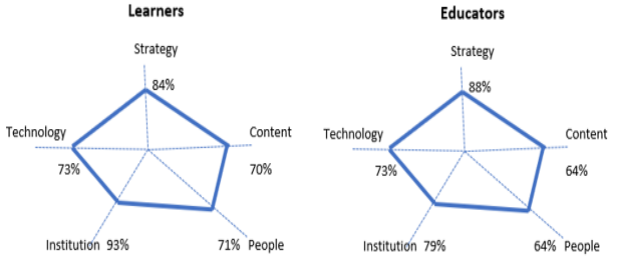


Figure 8. STIPC=STOPE comparison assessment results

Figure 8 shows overall assessment results of all the frameworks from the proposed model whereby learners' frameworks are as follow: institution is leading with 93 percent, followed by strategy with 84 percent, then technology with 73 percent and content is 70 percent while people is 71 percent. The educators' overall results are low compared to leaners 'results. The results are as follow: the strategy is leading with 88 percent, followed by institution with 79 percent, then technology with 73 percent and content together with people are both on 64 percent.

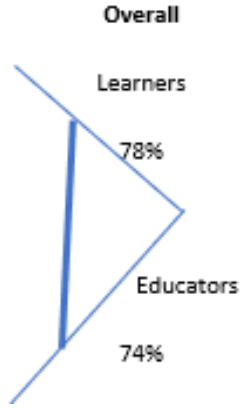


Figure 9. Overall STIPC = STOPE e-learning readiness indicator

Figure 9 shows the overall stope assessment that were derived from Table IV. The learners are on 78 percent which is higher level than the 74 percent of educators. This clearly shows that the educators have negative attitude towards e-learning. In terms of technological skills, it is clear that the learners are familiar with technology than the educators due to their age.

TABLE V. TABLE REPRESENTATION OF EDUCATORS AND LEARNERS' PROFILE

Educators				Learners		
GENDER	**Number**	**%**		**GENDER**	**Number**	**%**
Male	21	44%		Male	61	48%
Female	27	56%		Female	67	52%
Total	48	100%		Total	128	100%
AGE				AGE		
≤30	5	10%		≤13	11	9%
31-45	17	35%		14-17	96	75%
≥46	26	54%		≥18	21	16%
Total	48	100%		Total	128	100%
SOCIAL MEDIA				SOCIAL MEDIA		
Twitter	11	23%		Twitter	68	53%
Facebook	28	58%		Facebook	90	70%
LinkedIn	2	4%		LinkedIn	3	2%
Whatsapp	39	81%		Whatsapp	108	84%
Instagram	4	8%		Instagram	77	60%

As mentioned in methodology, the educators' population in this paper consisted of 48 educators of which 21 are male and 27 are female. The learners' population consisted of 128 learners of which 61 are male and 67 are female. The population age group is summarised in table 4 above.

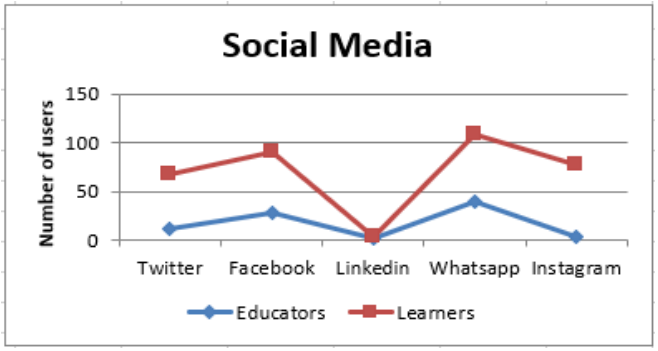


Figure 10. Social media results graph

As per figure 10 above, the learners are millennials and technologically skilled in terms of social media while educators are below. The highest used social media in both learners and educators is whatsapp, followed by facebook, instagram then twitter.

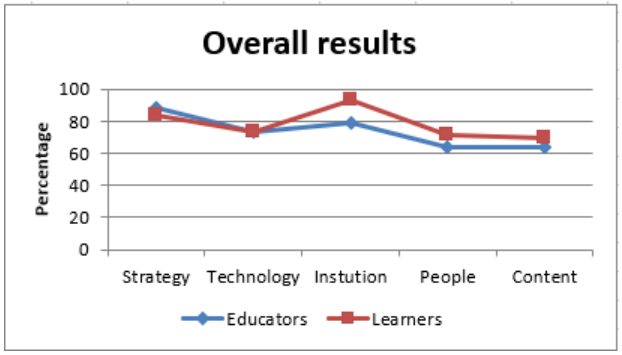


Figure 11. Overall comparative results graph

The results in figure 11 are derived from figure 8. The figure is showing the overall results of all the variables between the educators and the learners.

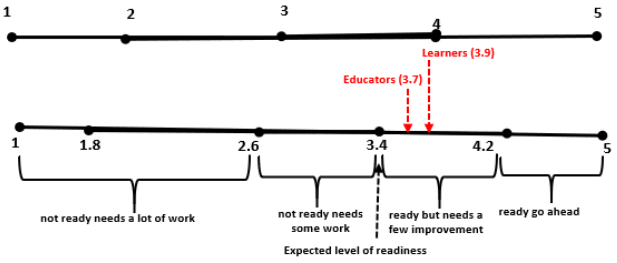


Figure 12. Learners and educators' e-learning readiness level in on a likert scale

The expected level of readiness is 3.4 given by [17, 18]. Using the five-point likert scale as shown in figure 28, the

readiness levels are 3.7 and 3.9 for educators and learners respectively. But when using the weighted average, the readiness levels is 2.8 and 3.1 respectively.

The educators' readiness level ranges from 2.8 to 3.7, with a variance of 0.9 or 24%. The learners' readiness level ranges from 3.1 to 3.9, with a variance of 0.8 or 26%.

IV. CONCLUSION

In this research, e-learning readiness was the main concern. For e-learning benefits to be obtained, it is important to understand the level of readiness. It is remarkable to see the readiness of the learners being higher than what the educators are. It is clear that the learners are millennials and technologically skilled while educators are unfamiliar with technology because of age and attitude. Learners levels of readiness in terms of content (70%), people (71%), strategy (84%) and institution (93%) are all higher than the levels of educators. Then technology is on the same level of 73% on both the learners and educators. Both content and people levels for educators are on 64%. The educators need to be educated about e-learning to improve their readiness level. The educators' age and attitude issues need to be investigated more in future work to understand the educators' challenges towards e-learning.

REFERENCES

[1] M. Mawoyo, "ICT for Teaching and Learing in South Africa," *ict Development Associates ltd,* pp. 150-169, 6 June 2011.

[2] I. F. Maulida and J. Lo, "E-Learning Readiness in Senior High School in Banda Aceh," *Journal of Information Technology and Applications,* vol. 7, No. 4, pp. 122-132, 2013.

[3] C. Graham and J. Bonk, *The handbook of Blended Learning*. San Francisco: Pfeiffer, 2006.

[4] T. Bates, R. S. i. D. Education, Ed. *Technology, e-learning and distance education.*, 2nd Edition ed. London ; New York. : Routledge, 2005, p. 260.

[5] M. Rosenberg, "E- Learning; Strategies for Delivering Knowledge in the Digital Age. ," in *E-Learning and Consulting Philips Medical Systems North America*, P. E. Sanderson, Ed. USA: Elsevier Science Inc., 2001, pp. 185-188.

[6] S. Chappuis, R. J. Stiggins, J. Arter, and J. Chappuis, *Assessment for Learning: An Action Guide for School Leaders* (Assessment Training Institute, Inc.). Pearson; 2 edition (March 11, 2009), 2005, p. 288.

[7] V. Arkorful and N. Abaidoo, "The role of e-learning, the advantages and disadvantages of its adoption in Higher Education.," *International Journal of Education and Research,* vol. Vol. 2, December 2014.

[8] H. A. Oketch, J. M. Njihia, and A. N. Wausi, "E-learning readiness assessment model in Kenyas' Higher Education Institutions: A case study of university of Nairobi," *International Journal of Scientific Knowledge Computing and Information Technology,* vol. 5, no. 6, 2014.

[9] C. T. Lopes, "Evaluating e-learning readiness in a health sciences higher education institution," 2008.

[10] G. Omoda and J. Lubega, "E-learning Readiness Assessment Model: A case Study of Higher Institutions of Learning in Uganda Hybrid Learning," vol. 200-211, 2011.

[11] A. A.-F. Alshaher, "The McKinsey 7S Model framework for e-learning system readiness assessment," *International Journal of Advances in Engineering & Technology,* vol. 6, no. 5, pp. 1948-1966, Nov. 2013.

[12] A. Mercado, "Readiness Assessment Tool for An eLearning Environment Implementation," presented at the Fifth International Conference on eLearning for Knowledge-Based Society, Bangkok, Thailand, 2008.

[13] İ. Soydal, G. Alir, and Y. Ünal, "Are Turkish Universities Ready for Elearning: A case of Hacettepe University Faculty of Letters," 2012.

[14] D. o. Education, "Education Annual Report 2016/2017 Financial Year," in "Province of Gauteng," 2017, vol. 5.

[15] S. H. Bakry, A. Khalid, and A. Abdulmohsen, "STOPE-3. based Approach for e-Readiness Assessment Case Studies: Department of Electrical Engineering," *International Journal of Network Management,* vol. 18, pp. 65-75, 2007.

[16] K. Al-Osaimi, A. Alheraish, and S. H. Bakry, "STOPE-based approach for e-readiness assessment case studies," *International Journal of Network Management,* no. 18, pp. 65-75, 2008.

[17] S. Nisperos, "Assessing the E-Learning Readiness of Selected Sudanese Universities," *Asian Journal of Management Sciences & Education,* vol. 3(4), October 2014.

[18] O. Ouma, M. F. Awuor, and B. Kyambo, "E-learning readiness in public secondary school in Kenya," *European Journal of Open, Distance and e-Learning,* vol. 16 / No. 2, 2013.

A New Tree Routing Protocol for ZigBee Healthcare Monitoring Systems

Bongisizwe Erasmus Buthelezi
Department of Information Technology
Tshwane University of Technology
Pretoria, South Africa
bongisizwe.erasmus@outlook.com

Maredi I. Mphahlele
Department of Information Technology
Tshwane University of Technology
Pretoria, South Africa
MphahleleMI@tut.ac.za

Deon Du Plessis
Department of Information Technology
Tshwane University of Technology
Pretoria, South Africa
DuPlessisDP@tut.ac.za

Solly Maswikaneng
Department of Information Technology
Tshwane University of Technology
Pretoria, South Africa
MaswikanengPS@tut.ac.za

Topside E. Mathonsi
Department of Information Technology
Tshwane University of Technology
Pretoria, South Africa
MathonsiTE@tut.ac.za

***Abstract*—ZigBee is an industrial standard for personal area network (PAN) developed for low power, cost and rate for wireless radio communications. Lately, ZigBee implementation in Healthcare Monitoring Systems (HMS) is exponentially increasing. ZigBee specification proposed that a number of wireless sensor nodes can be connected through a tree hierarchal topology in a ZigBee network. In Health sector, HMSs implement body sensor nodes that are capable of collecting human physiological data such as Heart Rate signals. However, routing in tree topology needs every sensor node to transmit data to a central node called coordinator. Parent-child is the mechanism that is used to forward packets to the coordinator in the original ZigBee Tree Routing (ZTR) scheme. However, parent-child mechanism is not suitable to transmit sensitive data such as heart rate pulses. Parent-child mechanism suffers from end to end delay problem which is a disadvantage for physiological signals as they require emergency attention. This paper is proposing a solution that attempts to solve this problem by implementing New Tree Routing Protocol (NTRP) algorithm. NTRP scheme is evaluated on NS-2 simulator tool and its simulation results show improvement of average end to end delay of 0.3-1% while the packet delivery ratio is 5-11.6%.**

***Keywords*—ZigBee, Parent-child, end to end delay, Heart Rate Pulse, Tree routing, NTRP**

I. INTRODUCTION

Wireless Sensor Network (WSN) is a collection of sensory nodes that assists in sensing and monitoring environments such as healthcare, military, homes and other fields [1]. In health care, sensor nodes can sense health related data parameters such as Heart Rate Pulse, High Blood pressure and other cardio vascular diseases through Health Monitoring Systems. According to the report released by the World Heart Federation [2] in their 10th anniversary of World Heart Day, heart disease is one of the critical diseases that accounts for about 17 million deaths a year worldwide. Elderly people are the victims of heart attacks due to their ageing related challenges. Some do not have health insurances and medical aid schemes or they may have inadequate access to it.

This creates a challenge hence they might need more assistance at home or at the healthcare facility such as traditional old age home. To address this challenge, a number of studies have been conducted based on Ambient Intelligence and Assisted Living to develop solutions that strengthen and support human physiology [3]. Technological based solutions such as Healthcare Monitoring Systems (HMSs) reduce costs and provide life support to the needy. This paper is trying to solve end to end delay of Heart rate data sent over ZigBee network in HMSs to assist elderly people in need of quick assistance with their heart disease.

With the assistance of Healthcare Monitoring Systems, elderly people's health can be monitored to identify abnormal physiological patterns. If a condition such as abnormal heart rate pulse occurs, the system can send a notification such as an alarm or short message (sms) to report the condition to the person in charge. HMSs technologies assist in real-time sensing and analysing medical data and they exchange analysed data between the elderly person and the doctor or nurse in charge. The popularity of HMSs is heightened by their new methodology of remotely monitoring elderly people in Ambient Assisted Living (AAL) environments such as homes, old age home etc.

This paper adopts a 3-Tier Healthcare Monitoring System architecture [4], but the focus is on a ZigBee communication routing protocol in Tier-2. As depicted in Fig. 1 below, the architecture is composed of Body Area Network, ZigBee Wireless Personal Area Network, and a Wide Area Network.

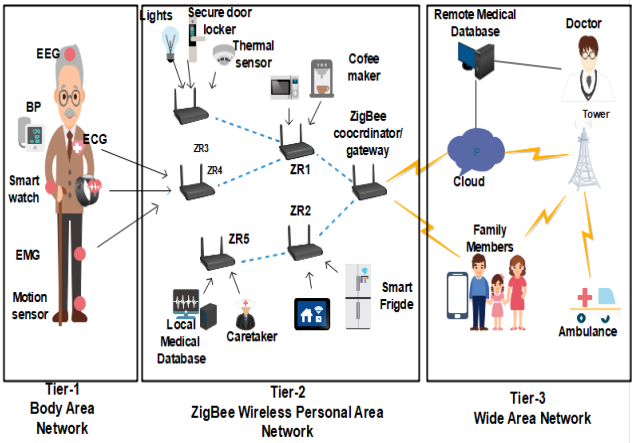


Fig. 1. Healthcare Monitoring System architecture.

The rest of this paper is organized as follows: Section II, gives the background of ZigBee network, devices, topologies and protocols associated. Section III, provides the related literature. Section IV, presents the proposed algorithm as a solution to the problem. Section V presents the performance evaluation for a New Tree Routing Protocol algorithm. Conclusion and future work are presented in Section VI.

II. Background And The Challenge Facing ZigBee Communication Protocol For Rate-Wireless Personal Area Network

In this paper, Healthcare Monitoring Systems are classified as Traditional Healthcare Monitoring Systems and Smart Healthcare Monitoring Systems. Smart Healthcare Monitoring Systems are further broken down into three categories;

- Remote Healthcare Monitoring Systems [5] are designed to send and receive data from a remote location where distance separates patients from healthcare professionals.
- Mobile Healthcare Monitoring Systems [6] are designed to capitalize on mobile devices such as smart phones to relay data between healthcare professionals and patients.
- Wearable Healthcare Monitoring Systems [7] focus on utilizing wearable devices that are embedded on clothes as sensor nodes that forms body sensor network. However the chosen category for this paper is remote healthcare monitoring systems with the main interest of enhancing ZigBee wireless communication technology.

A. ZigBee Devices and topologies

ZigBee network is composed of three types of devices namely; the coordinator, routers, and the end devices. These devices can be connected into several topologies such as star, peer to peer, cluster tree and mesh topology as depicted on the below Fig. 2.

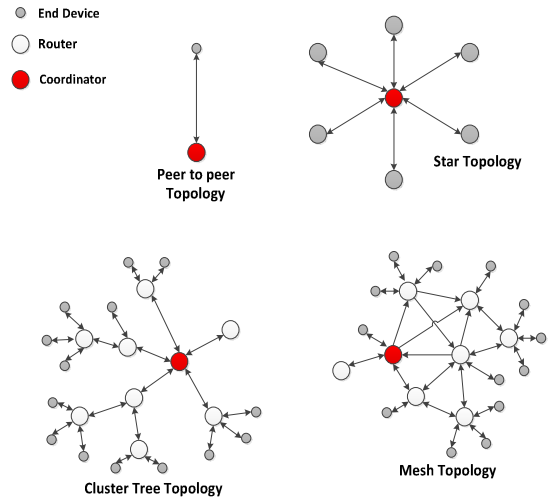


Fig. 2. Classification of ZigBee Topologies

A Coordinator is the most capable ZigBee device that controls the network and it also acts as a gateway router to the outside networks. To form a new network, the coordinator select a single channel and network ID (PAN ID) and then other devices can join the network. In Healthcare Monitoring Systems routers are responsible for end to end data communication. These devices are capable of selecting the best route a packet can take in order to avoid delay, packet loss, and traffic jam. In Healthcare Monitoring Systems end devices are sensor devices that directly collect health data from a human body and send it to ZigBee Router devices. Wearable sensors are attached to a patient's cloth and body sensors implemented under a human skin to collect cardio vascular data signals.

Wireless communication protocols such as Bluetooth, Wi-Fi, and ZigBee transmit data in a short range, therefore in Healthcare Monitoring Systems these protocols are chosen based on their ability to reduce cost, low energy consumption, speed, and security [8]. Unlike Wi-Fi and Bluetooth, ZigBee is an IEEE 802.15.4 standard that is suitable for implementation in HMSs. Based on ZigBee specification [9] that was released in 2004, ZigBee network is suitable for Personal Area Networks such as Home Area Network (HAN) as it can reduce cost through its ability to consume low energy. This is also supported by a survey study conducted [10], where communication technologies were studied and compared to identify their strengths and shortcomings. Mainly, ZigBee communication technology is used to assist elderly people in Ambient Assisted Living environments [11]. The only challenge with ZigBee network is its ZigBee Tree Routing protocol that causes a high amount of end to end delay of packets in a network.

B. ZigBee Tree Routing Protocol

In ZigBee network, a coordinator is a router device that acts as a parent for all routers, therefore, all nodes send their data to the coordinator. End devices cannot be parents, but they can be only child because they lack routing capability. A child device is connected to a parent device and messages intended for a child can be routed through its parent. ZigBee coordinator offers each potential parent with sub-block of network addresses, and a parent node (router) assigns addresses to its children (end devices). ZigBee wireless sensor network uses ZigBee Tree Routing (ZTR) protocol to find the best path to the destination. ZTR forwards packets based on a parent-child mechanism. This mechanism allows packets to pass through many hops before they reach their destination.

C. Problem with the ZigBee Tree routing protocol

Parent-child relationship is the main shortcoming of ZTR protocol. It allow packets to be routed from the source node along the tree topology towards the destination node even if when the destination node located nearby [12]. In a parent-child relationship mechanism, the coordinator act as a middle node used by all other nodes to communicate on a network. Once the coordinator receives data, it forwards it down the branches of the tree towards the destination node. This creates end to end delay as packets traverses through a number of hops before reaching the destination node. This is a challenge in Healthcare Monitoring Systems as they carry vital health data such as Heart Rate signals that require emergency treatment.

To address this challenge, this paper proposes a New Tree Routing Protocol (NTRP) that minimizes end to end delay by finding shortest route to the destination. NTRP solution minimizes the number of hops between a source and a destination node. The proposed solution implements clustering mechanism and neighbor table scheme in its algorithm. Furthermore, NTRP merges the classical ZTR algorithm with Kruskal's Minimum Spanning Tree to minimize end to end delay. The details of the proposed solution are discussed in section IV. The next section discusses related studies that have attempted to solve a parent child problem.

III. Related Works

ZigBee/IEEE 802.5.4 is a Low-Power standard for wireless sensor network that uses a classical ZigBee Tree Routing algorithm for packet routing. However, as it is discussed in the previous section, ZigBee Tree Routing algorithm have shortcoming of high end to end delay. This study is proposing a solution to minimize the end to end delay created by the parent child mechanism in a ZigBee Tree Routing algorithm.

In response to this problem, a number of studies attempted to provide different solutions. However as the number of new devices are added on the network the problem persist. A study [13] proposed an improved Tree Routing algorithm that utilizes the information of the neighbors within two hops. Their proposed algorithm allows nodes to use the distributed address assignment mechanism (DAAM) to calculate their depth. For a node to receive the 2-hop neighbor information easily, routing nodes broadcast beacon messages on a network. However, broadcasting messages increase the cost overheard and end to end delay.

A Shortcut Tree Routing (STR) is proposed to provide a near optimal routing path [14]. The SRT uses 1-hop neighbor information to maintain the advantages of the classical ZTR such as minimal or no route discovery overhead and low memory consumption. To achieve this, a set of predefined paths can be used for forwarding packets instead of calculating route dynamically. Each source node forwards packets to a neighbor with the smallest remaining hops in its neighbor table.

A Selfishness Aware Dynamic Source Routing (SA-DSR) [15] is proposed to improve a reactive routing protocol called Dynamic Source Routing (DSR). SA-DSR resolves energy consumption problem in ZigBee, but end to end delay problem remains a challenge. Instead of using neighbor tables, SA-DSR utilizes routing tables which leads to end to end delay problem. To address end to end to delay and energy consumption, an Enhanced Self Configuration (ESC) scheme is proposed [16] ESC minimizes the number of active ZigBee routers by turning off other routers. However, this scheme compromises packet delivery ratio while avoiding a problem of orphan propagation.

Another study conducted [17] uses multipath extension to forwards packets. Z-MHTR is a ZigBee Multipath Hierarchical Tree Routing proposed to enhance the overall performance of the classical ZTR protocol. Multipath routing solution is a disjoint multipath routing extension for the ZigBee protocol based on cluster tree mechanism. Cluster Tree structure based on a parent child established relationship helps Z-MHTR to make best forwarding decision. However, Z-MHTR has a shortcoming of end to end delay because of parent child relationship which has a number of hops between the source and the destination node. The study [18] proposed cluster tree to reduce energy consumption in order to extend the life time of the ZigBee network. The proposed solution also finds better paths to forward data packets.

Velocity Energy-efficient and Link-aware Cluster-Tree (VELCT) is the mechanism for data collection in WSN [19]. VELCT is designed to minimize energy exploitation, reduce traffic in cluster head and minimizes end to end connection. VELCT algorithm constructs a tree structure that reduces energy consumption of the cluster head and avoids frequent cluster formation. NS-2 simulator is used to evaluate the proposed algorithm. However, this solution takes a lot of memory space which compromises the overall network performance of the network. Therefore, based on the above studies it is concluded that there is no single solution to WSNs problems. However, combining some of the mentioned solutions can increase the performance of the network.

IV. Proposed Solution

A New Tree Routing Protocol (NTRP) merges ZigBee Tree Routing algorithm with Kruskal's Minimum Spanning Tree algorithm [20]. The proposed NTRP minimizes end to end delay of packets sent over the ZigBee network. Neighbor table and clustering mechanism are defined in the ZigBee specification [21]. Therefore, NTRP is designed by grouping devices into clusters based on their functionality such as grouping body sensors in a single cluster and home appliances in a single cluster. As seen in Fig. 3 below, in this study ZigBee network is divided into four clusters namely; Home appliances, CCV monitoring sensors, body sensors, and devices such as servers and computers. Each cluster has one cluster head that communicates directly with other neighbor cluster heads using neighbor tables. This proposed mesh topology minimizes delay to the destination node.

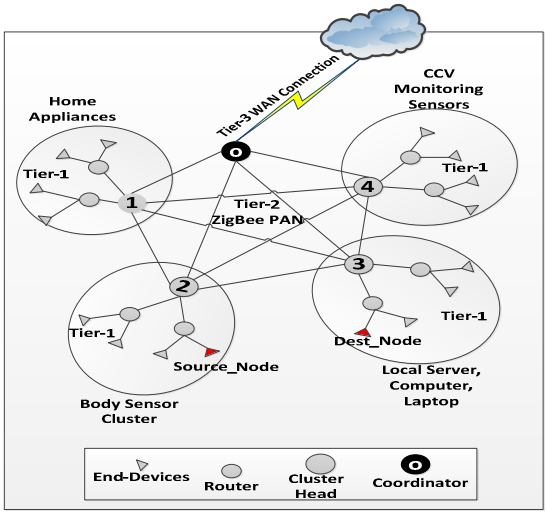


Fig. 3. Proposed Mesh topologies divided into clusters

The proposed algorithm below first calculates the heart rate threshold level received from body sensors. Threshold level for heart is calculated based on the maximum and minimum levels of heart pulse signals. Once the threshold value is calculated, it used to design the heart rate pulse algorithm. The conditions of the algorithm are as follows:

- If the maximum heart rate pulse is in range of 60 – 100 Bpm, the condition for a patient is normal.
- If the maximum heart rate pulse is lower than 60 Bpm, the condition for a patient is abnormal and it is called Bradycardia.
- If the maximum heart rate pulse is above 100 Bpm, the condition for a patient is abnormal and the condition is called Tachycardia.

If the condition is abnormal an alert is sent to the doctor or a nurse to request emergence assistance.

NTRP ROUTING PROTOCOL FOR END TO END DELAY MINIMISATION

```
1.#include<iostream>
2.#include<vector>
3.#include<utility>
4.#include<algorithm>
5.usingnamespacestd;
6.constintMAX=1e4+5;
7.intNd_n[MAX],nodes,Src_Node,Dest_Node,HR_pulse,NTR_Addr
8. Max_Dept,Dest_Node=Find_max(Addr(Dest_Node),0);
9 {  if (min_HR<60 && max_HR<100);
10     Status_condition=normal;
11      elseif (min_HR<50 && max_HR<60);
12          Status_condition= abnormal;
13      elseif (min_HR<100 && max_HR<100);
14          Status_condition= abnormal;
15      endif  }
16 {  if Nd_n is an FFD && Dept(Y)<rL_m –1 && Nd_n(Y)<rR_m then
17     Nd_p(Y)←Nd_p(Y)+1    //accomodating node as a ZigBee parent
                                router, allocate the following address
18.    Addr(Y)=rC_skip(Dep(Y))×(Nd_p(Y)–1)+1;
19.    elseif Dep(Y)<rL_m –1 && Nd_c(Y).rC_m then;
20.        Nd_c(Y)←Nd_c(Y)+1    //accomodating node as a child node
                                allocate the following address
21.        Addr(Y)=rC_skip Nd_p(Y))×rR_m +Nd_c(Y);
22     else    //node cannot be accomodate
23.   rC_Skip(d) = (1+rC_m –rR_m rR_m^(rL_m–d–1)) / (1–R_m)  then;
24 }  void initialize()
25.   {  for int x=0;x<MAX;++x);
26.        Nd_n[x]=x;
27.    }  int root(int i);
28.      {  while(Nd_n[i] !=i);
29.         {  Nd_n[i]=Nd_n[Nd_n[i]];
30.          i=Nd_n[i];
31.          }
32.           return i;
33.    }
```

A. Routing Rules for NTRP Algorithm

The next step for the NTRP algorithm is to execute routing rules. NTRP use cluster head ID to find neighbor cluster heads. With the assistance of Kruskal's Minimum Spanning Tree, a route with minimal weight that can reduce end to end delay is selected. All cluster heads sends hello packet messages to collect information about the surrounding neighbors. Source node sends Heart rate threshold value to the cluster head. Cluster head forward a packet with heart rate signal to its neighbor cluster heads using neighbor table. If the destination node is not found in local clusters, the coordinator forwards the packet outside the network to any third party connected to the ZigBee network. Information about neighbor nodes (cluster heads) is stored in a cluster head's neighbor table. In the above NTRP algorithm, if the message is sent to node (n) then n simply accepts the packet otherwise it performs the following operations:

- If the receiving node is a neighbor of N_d, then N_d directly send the packet to the destination node.
- If the next hop is the cluster of N_d then N_d is in the same cluster with Cl_n.
- If N_{src} is an ancestor cluster of Cl_n such that $Cl_n < dest \le Cl_n + rC_m - 1) \times rC_{skip}(d) + 1$, then N_d looks if it has a neighbour N_d that meets the conditions of $Cl_n \le N_{src} \le Cl_n + (rC_m - 1) \times rC_{skip}(d+1) + 1.$

V. Simulations And Experimental Results

Simulations were conducted to evaluate the performance of the proposed NTRP protocol in comparison to STR and VELCT protocols. STR and VELCT were proposed to improve the original ZTR; therefore, in this paper it is assumed that both STR and VELCT have better performance than the original ZTR. End to end delay and packet delivery ratio were utilized to compare NTRP, STR and VELCT routing protocols. In this paper, NS-2 simulation tool is used as a test platform for the proposed NTRP protocol. The below Table I, depicts the parameters used for the simulations performed in this study.

TABLE I. Simulation Parameters

Simulation tool	NS-2
Number of nodes	40 - 240
Simulation Area	200m * 200m
Propagation Model	Two-Ray
PHY/MAC protocol	IEEE 802.15.4
Queuing	Priority Queue
Network Protocol	NTRP/STR/VELCT
Simulation Time	300s
Packet Type	CBR
Nodes Deployment	Random
Max. Tx Range	50m
Packet Size	64 Bytes

A ZigBee coordinator node is deployed in a fixed position while other ZigBee nodes such as routers and end devices are randomly deployed.

A. End-to-End Delay(E2E)

In this paper, end to end delay is defined as the time taken when a packet leaves the source node and the time a packet reaches its destination node. Fig. 4 below depicts the average end to end delay for all three routing protocols that were compared in this study. It is noted that from the simulation results, the end to end delay for all three protocols increases as more nodes are deployed in the simulation test. This is because the topology becomes more complicated when more nodes are deployed.

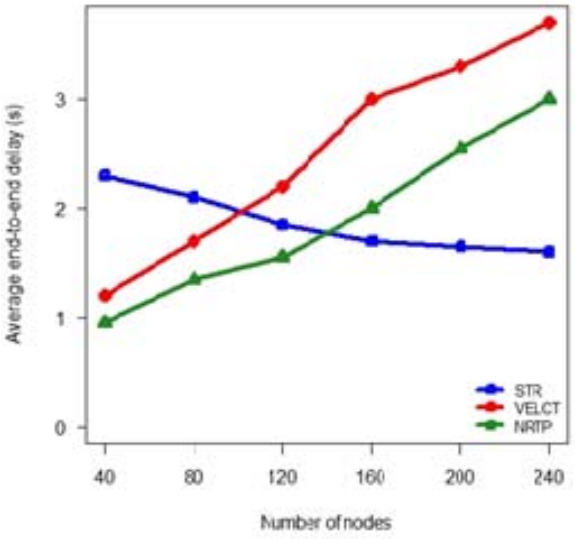


Fig. 4. Average End to End Delay simulation results

However, it is observed that the average end to end delay for the proposed NTRP is less than that of STR and VELCT. This is because NTRP eliminate a number of hop count by using neighbor tables and it also send packets to other clusters via cluster heads in a point to point fashion. Thus NTRP routing protocol reduces 0.3% to 1% of average end to end delay compared to STR and VELCT as seen in TABLE II below. However, STR outperforms the proposed NTRP when the number of nodes is above 120.

TABLE II. OVERALL PERFORMANCE COMPARISON OF END TO END DELAY.

Number of Nodes	STR	VELCT	NTRP
40	2.3%	1.2%	0.95%
80	2.1%	1.7%	1.35%
120	1.85%	2.2%	1.55%
160	1.7%	3%	2.0%
200	1.65%	3.3%	2.55%
240	1.6%	3.7%	3.0%
Tot E2E Delay	**1.8%**	**2.5%**	**1.5%**

It is observed that NTRP has lower total end to end delay (Tot E2E Delay) compared to other protocols. NTRP has Tot E2E Delay of 1.5%, STR has an average of 1.8% and VELCT has an average of 2.5%. The results show that when more nodes are added, end to end delay increases in all three algorithms.

B. Packet Delivery Ratio

Packet delivery ratio in this paper is defined as the ratio of data packets received by the destination node in comparison to the total number of packets generated by the source node. The simulation results for the packet delivery ratio for all three routing protocols is depicted in Fig. 5 below.

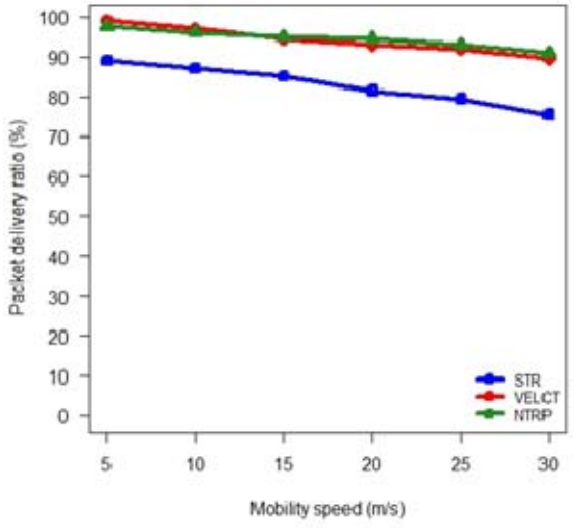


Fig. 5. Packet Delivery Ratio simulation results

From the above results, it can be observed that the packet delivery ratio for all three routing protocol decreases as the number of nodes increases. This is because there is high number of collisions that occur at the lower layer nodes in a cluster-tree topology. In response to this challenge, NTRP utilizes neighbor cluster heads as its next hop. In this way, NTRP decreases the collision of packets in lower layer nodes and creates shorter paths towards the destination. As a result, NTRP achieves higher packet delivery ratio of 5% - 11.6% compared to other routing protocols as depicted in TABLE III below.

TABLE III. OVERALL PERFORMANCE COMPARISON FOR PACKET DELIVERY RATIO

Mobility Speed	STR	VELCT	NTRP
5	89%	99%	97.5%
10	87%	97%	96%
15	85%	94%	95%
20	81.5%	93%	94%
25	79.5%	92%	93%
30	75.5%	89.5%	91%
Tot PDR	**82.9%**	**94%**	**94.5%**

Table III above depicts that the average percentage of packet delivery ratio is 94.5% for NTRP, 94% for VELCT and 82.9% for STR.

VI. Conclusion and Future Work

In this paper, NTRP algorithm is proposed to minimise the end to end delay for a ZigBee network in Healthcare Monitoring Systems. The proposed technique merges the original ZigBee Tree Routing algorithm and Kruskal's algorithm to find shortest routes from source to destination node. Clustering technique and neighbour tables are additional features to complement the performance of the proposed solution. NS-2 simulation tool is used to evaluate the proposed NTRP where average end to end delay is 0.3 – 1% while the packet delivery ratio is 5-11.6%. Therefore, NTRP solution promises to provide quick detection of heart related problems for elderly in Ambient Assisted Living environments. Future scope for this paper will consider implementation of secure end to end encryption of medical data sent over Healthcare Monitoring Systems (HMSs). Blockchain technology could be used to secure medical data. This is because HMSs rely on several technologies that can pose security threats such as modifying medical data and theft of medical data.

Acknowledgment

This work is part of MTech dissertation, the authors would like to thank Tshwane University of Technology with their support.

References

[1] T. D. P. Mendes, R. Godina, E. M. G. Rodrigues, J. C. O. Matias and J. P. S. Catalão, "Smart Home Communication Technologies and Applications: Wireless Protocol Assessment for Home Area Network," *Energies,* vol. 8, no. 7, pp. 7279-7311, 2015.

[2] B. Remenyi, J. Carapetis, R. Wyber, K. Taubert and B. M. Mayosi, Heart State of the Heart: Cardiovascular Disease Report, Switzerland: World Heart Federation, 2012.

[3] A. P. Abidoye, N. A. Azeez, A. O. Adesina, K. K. Agbele and H. O. Nyongesa, "Using Wearable Sensors for Remote Healthcare Monitoring System," *Journal of Sensor Technology,* vol. 1, no. 2, pp. 22-28, 2011.

[4] C. Dobre, G. Mastorakis, N. Garcia, R. I. Goleva and G. Mastorakis, "Ambient Assisted Living and Enhanced Living Environments: Principles, Technologies and Control," in *Introduction to the AAL and ELE*, Butterworth-Heinemann, 2016, pp. 1-16.

[5] K. Guan, M. Shao and S. Wu, "A Remote Health Monitoring System for the elderly based on smart home gateway.," *Journal of Healthcare Engineering,* vol. 2017, pp. 1-9, 2017.

[6] V. Wahane and P. Ingole, "Interactive Mobile Health Monitoring System," *International Journal of Advanced Computer Science and Applications (IJACSA),* vol. 8, no. 4, pp. 304 - 310, 2017.

[7] M. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang and C.-H. Youn, "Wearable 2.0: Enabling human-cloud integration in next generation healthcare systems," *IEEE Communications Magazine,* vol. 55, no. 1, pp. 54-61, 2017.

[8] ZigBee, "ZigBee Specification Document 053474r20," ZigBee Standards Organization, San Ramon, 2012.

[9] B. Thaduangta, P. Choomjit, S. Mongkolveswith, U. Supasitthimethee, S. Funilkul and T. Triyason, "Smart Healthcare Basic health check-up and monitoring system for elderly," in *2016 International Computer Science and Engineering Conference (ICSEC)*, Chiang Mai, 2016.

[10] B. E. Buthelezi, T. E. Mathonsi, S. Maswikaneng and M. Mphahlele, "Routing Schemes for ZigBee Low-Rate Power Personal Area Network: A Survey," in *11th International Conference on Data Mining, Computers, Communication and Industrial Applications (DMCCIA-2017)*, Kuala Lumpur, 2017.

[11] R. Elankavi, R. Kalaiprasath and R. Udayakumar, "Wireless ZigBee Network Cluster-Capacity Calculation and Secure Data Conveyance Using Indegree," *International Journal on Smart Sensing and Intelligent Systems Special Issue,* vol. 1, pp. 175-185, 2017.

[12] Y. Yao, Z. Zhang, Z. Ren and S. Shi, "Neighbor information assisted tree routing for zigbee networks," in *International Conference on Computer Science and Electronics Engineering*, 2012.

[13] T. Kim, S. H. Kim, J. Yang, S.-e. Yoo and D. Kim, "Neighbor Table Based Shortcut Tree Routing in ZigBee Wireless Networks," *IEEE Transactions on Parallel and Distributed Systems,* vol. 25, no. 3, pp. 706 - 716, 2014.

[14] D. G. Kampitaki and A. A. Economides , "Novel routing protocol for Mobile Ad Hoc Networks with selfish and altruistic nodes," in *International Conference on Telecommunications and Multimedia (TEMU)*, 2016.

[15] A. J. Krishnan and G. S. Binu , "Energy efficient tree construction for ZigBee router network," in *International Conference on Communication and Electronics Systems (ICCES)*, 2017.

[16] Z. Bidai and M. Maimour, "Interference-aware multipath routing protocol for video transmission over ZigBee wireless sensor networks," in *International Conference on Multimedia Computing and Systems (ICMCS)*, 2014.

[17] R. Velmani and B. Kaarthick, "An efficient cluster-tree based data collection scheme for large mobile wireless sensor networks," *IEEE sensors journal,* vol. 15, no. 4, pp. 2377-2390, 2015.

[18] B. Munier, M. Aleem, M. A. Islam, M. A. Iqbal and W. Mehmood, "A Fast Implementation of Minimum Spanning Tree Method and Applying it to Kruskal's and Prim's Algorithms," *Sukkur IBA Journal of Computing and Mathematical Sciences,* vol. 1, no. 1, pp. 58-66, 2017.

[19] F. Meng, B. Jiang and C. Wang, "An improvement of ZigBee cluster tree routing protocol," in *International Conference on Advances in Mechanics Engineering*, 2012.

Performance Analysis of Distribution Networks under high Penetration of Photovoltaics in Mauritius

Waseem Ahmad Nosimohomed
School of Engineering and Physical Sciences
Heriot-Watt University
Edinburgh, UK
wn50@hw.ac.uk

***Abstract*—Traditionally, for a radial low-voltage (LV) distribution network serving residential loads, it was simpler for the electric utility to predict the demand and adjust the generation from power plants accordingly, thus offering a high reliability of supply. However, with the increasing penetration of photovoltaics (PV) on roof-tops of residential houses, it has become difficult to ensure the same reliability without the performance of the distribution network being affected. At times of low loads in residential houses, PV systems will inject power upstream of the host LV network and affect the behavior of the distribution network. In this paper, an LV distribution feeder in Mauritius has been modelled on DigSILENT Power Factory software to simulate how the network performance is affected under high PV penetration. Novel indices were developed for analyzing grid behavior in terms of maximum voltage variation from nominal value, average feeder loading, feeder losses, and transformer reserve. These performance indices were evaluated against three different voltage control strategies of PV inverters in order to choose the best one which will optimize the performance of the network. Furthermore, variations in the loads and PV generation over a 24-hour period were also considered in the performance analysis of the grid.**

Keywords—Distribution, indices, performance, photovoltaics, voltage.

I. INTRODUCTION

Many insular countries, like Mauritius, are highly dependent on imported fossil fuels for the generation of electricity to satisfy the load demand. Renewable energies, known as substitute technologies for conventional fossil-fuel technologies, were identified as the potential solution to mitigate the problems of increasing greenhouse gases and climate change. With the unit cost of photovoltaic cells having decreased radically as low as €0.40/Wp for high efficiency crystalline modules [1], this triggered a booming niche market leading to a widespread installation of photovoltaic systems in many countries.

Similarly, the target and policy of the Government of Mauritius is focused on increasing the amount of renewable energy to decrease its dependency on imported fossil fuels. One of its main target outlined in its Long Term Strategy Plan is to reach 35% of renewable in the energy mix of Mauritius by the year 2025, among which 2% is allocated to PV [2]. In order to achieve this, several strategies have been adopted by the Government, one of them being the introduction of schemes at the distribution level to promote and facilitate the installation of rooftop photovoltaics (PV) for different categories of customers. In 2018, a new scheme has been introduced in Mauritius, also known as the Home Solar Project, whereby 10,000 rooftop PV will be installed in different phases over a period of 5 years [3].

II. PROBLEM DEFINITION

Since the voltage profile decreases along the distribution feeder due to different line impedances and loads supplied, the onus is exclusively on the utility to keep the voltage within acceptable limits while supplying electricity to its customers. Solar photovoltaic (PV) facilitates the integration of renewable share as outlined by the policy of the Government of Mauritius. However, the connection of PVs to the distribution grid brings in some challenges: a mismatch between the production and the demand due to the stochastic generation of PV, discontinuity and bidirectional power flow which can affect the loading, voltage rise issues at the point of connection (PCC) and potential overloading of network equipment [4].

With the number of PV installations increasing significantly in Mauritius and bringing about changes in power flow and voltage and power flow as seen in Fig.1, it is obvious that the behavior of the distribution network will be affected [5] in terms of voltage deviation, reverse power flow, transformer and feeder loading capacity.

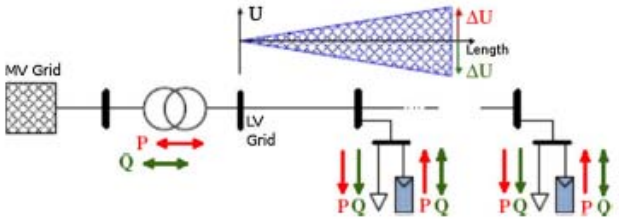


Fig. 1: Bidirectional power flow and voltage fluctuations in LV networks with PV integration [6]

Clearly, this imposes an operation constraint for the utility. For that reason, in many countries with PV systems, the utility's position is to apply the IEEE 1547 principle whereby any PV connected to the grid does not actively participate to regulate the distribution network voltage [7], which was deemed to be the best solution at that time.

With higher PV penetration levels, it can become cumbersome for the utility to control the voltage of the whole network independently without any participation from the PV, which is neither technically nor economically the best approach. With technology evolution, PV systems today have several modes of operation available, which are settable via the inverter, which allows them to inject reactive power in addition to real power.

Reactive power compensation by these modern inverters can be useful to regulate grid voltages by injection or absorption of reactive power [8]. There are two kinds of reactive power compensation, namely dynamic and static reactive support.

Dynamic reactive power support makes provision for grid stability in case of events like short-term sags. Inverters with this feature enabled allows the PV system to have "riding-through' capabilities. On the other hand, static reactive power support helps to keep the voltage within satisfactory levels. This feature is similar to the traditional method of operation, where reactive power is necessary for maintaining grid voltage stability.

The effectiveness of reactive power provision is based on the resistive R and reactive X shares on the local grid impedance (often expressed as X/R ratio). The higher the X/R ratio, the higher the effect of reactive power on the voltage magnitude. However, LV distribution networks have low X/R ratios, which hamper the technical effectiveness of reactive power provision on the local voltage magnitude [9]. Nevertheless, it still serves its purpose in LV networks by providing voltage support.

This paper is focused on assessing how the performance of the Mauritian low-voltage distribution network is affected under high penetration of PV and also determine the optimal mode of operation of all PV inverters to help assist in voltage support and regulation. This is investigated by using the static reactive power compensation with PV inverters providing voltage support under the following modes of operation: (i) Q(V) and (ii) Qmax in comparison with the PV inverters operating under the IEEE 1547 mode.

After an introduction to the operation modes of PV inverters considered in Section III, the methodology adopted to conduct this study is introduced in Section IV together with the definition of performance indicators. This is followed by illustration of the model developed on DigSILENT and running of load flows for different scenarios. The load flow results for each scenario are recorded, analyzed and discussed to reach the final conclusion in the Sections V, VI and VII respectively.

III. OPERATING MODES OF PV INVERTERS

Solar PV inverters are conventionally set to unity power factor. In other words, they maximize active power injection while reactive power injection is zero. Besides, if the PV inverters are permitted to inject reactive power too, they should then be able to provide the desired reactive power at maximum power generation. However, this can only be possible if the rating of the PV inverter is increased to provide for this additional output in active power [10], simply by oversizing the inverter as shown in Fig. 2. For the PV to operate under this mode, the PV inverter needs to be programmed to make use of the shaded grey area for supplying or absorbing reactive power depending on the network requirement. Nevertheless, oversizing the inverter has an added cost that has to be factored in, though the price differences are not big enough to hinder this decision [11].

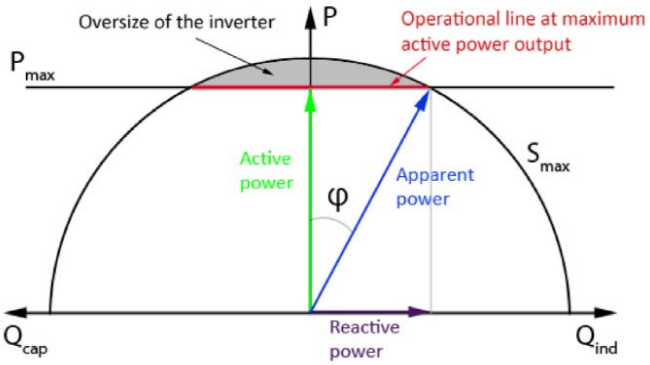


Fig. 2: Oversizing the inverter (Data Source: Gonzalez [11])

For this example, the added capacity factor, α, can be formulated as a proportion of the maximum active power generated, P_{pv}^{max}, where α varies between 0 to a user settable value (typically less than 1) as shown in Fig.3(a).

The reactive power capability obtained by oversizing the inverter can be used as per the different control approaches as shown in Fig. 3(b), (c) and (d) [12,13].

A. IEEE 1547 Control

Under this mode, the PV inverter the would not respond to voltage variations by changing reactive power production in a closed loop approach [14]. Therefore, the output of the PV inverter is only the active power, P_{pv} ranging from 0 to P_{pv}^{max}, as shown in Fig.3 (b), and the value depends on the power produced by the solar PV panels. This approach is adopted by many countries, but is not ideal since the voltage regulation constraint lies solely on the electric utility, which will be a real burden with higher PV penetration levels.

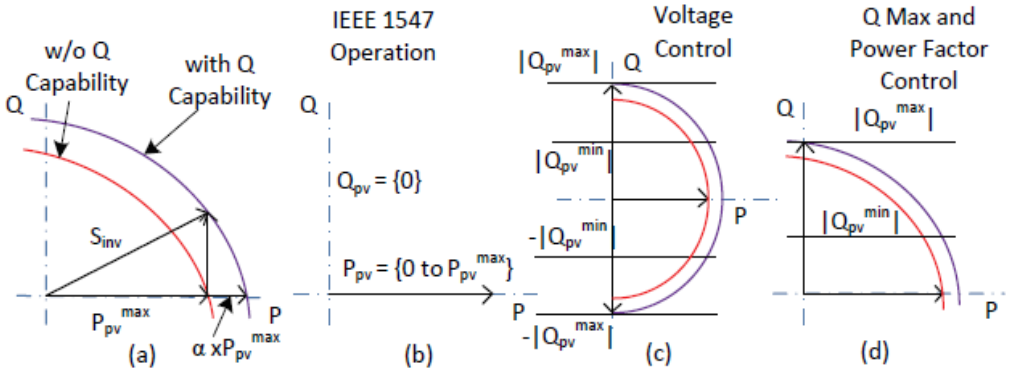


Fig. 3: PV inverter capability with different control strategies (a) Increase of PV inverter size for reactive capability addition (b) IEEE 1547 control (c) Voltage control (d) Q max and power factor control.

B. Q(V) or Voltage Control

This approach is the same as a conventional power generating unit that regulates the voltage by varying reactive power output within its reactive capability limits, with its active power output remaining in the same range of (0, P_{pv}^{max}). Depending on the voltage set-point, the PV inverter will determine the amount of reactive power injected or absorbed. The operating point will vary depending on this amount of reactive power within its capability limits, as shown in Fig. 3(c). Exceeding those limits will result in loss of voltage control and reactive power injected or absorbed becomes fixed at the limit value.

C. Qmax or Power Factor Control

Under this mode, the PV inverter allows the flow of reactive power upstream of the feeder in addition to active power in the same range of (0, P_{pv}^{max}). If the reactive power generated from the PV is higher than the reactive power demand at the PCC, then the resulting reactive power flow will in the reverse direction in the feeder. Solar PV inverters deliver maximum reactive power depending on their remaining capability after generating active power, as shown in Fig. 3(d). The amount of reactive power generation at any given active power generation, P_{pv}, and, additional capacity factor, α, can be defined by the following expression in Equation (1).

$$Q_{pv} = \sqrt{\{(1+\alpha)P_{pv}^{mqx}\}^2 - P_{pv}^2} \qquad (1)$$

IV. METHODOLOGY

The first step in order to assess how the behavior of a low-voltage distribution network is affected under high solar PV penetrations, is to develop performance indicators in relevance to the network parameters. These indicators will help evaluate the performance of individual components which constitute the network, namely the distribution transformer, LV feeder lines and consumer loads.

A. Development of Performance Indicators

In this study, we have developed network performance indicators in order to analyze the maximum voltage deviation, reserve capacity on the distribution transformer, average LV feeder loading and the corresponding losses with increasing levels of PV penetration.

1) Maximum Voltage Deviation Index (MVDI)

The voltage deviation is obtained by calculating the change in the nominal voltage with and without PV integration measured at the far end of the LV feeder line. Under a scenario with no PV connected to the system, the voltage may drop significantly along the feeder. Insertion of PV within the system will cause a rise in voltage in the grid. For any phase p, where p ∈ {a, b, c}, of a three-phase feeder, if the maximum voltage deviation is found at node k, then the maximum deviation index, MVDI can be calculated from the expression given in Equation 2,

$$MVDI^p = \frac{V_{nom} - V^{kp}}{V_{nom}} \qquad (2)$$

where V_{nom} is the nominal voltage and V_{kp} is the voltage at k th node of phase p. The change in voltage profile due to PV inclusion acts as voltage support as long as the voltage limit bandwidth is not exceeded. As per Equation (2), the value of MVDI will be negative when the node voltage at k exceeds the nominal voltage level.

2) Average Feeder Loading Index (AFLI)

Under normal operating scenarios, the loading on a distribution feeder should not exceed its rating capacity. Thus, using the average feeder loading is a good indicator to assess how much the feeder is overloaded. If the feeder line is divided into different segments, each segment will have different loading levels. An average feeder loading index, AFLI, can be defined based on the weighted average of percentage length of the feeder segments, as shown in Equation 3,

$$AFLI^p = \sum \left(\frac{l^p_{k(k+1)}}{L^p_f}\right)\left(\frac{\left|S^p_{k(k+1)}\right|}{C^p_{k(k+1)}}\right) \qquad (3)$$

where $l^p_{k(k+1)}$ is the length of the segment at phase p from the k-th node to the $(k+1)$-th node, L_f is the total feeder length, $S^p_{k(k+1)}$ is the apparent power flow in the segment under consideration and $C^p_{k(k+1)}$ is the rated apparent capacity of the segment. The average feeder loading will generally decrease with PV supplying the loads, but it may also increase due to reverse power flow due to large PV inclusions.

3) Feeder Loss-to-Load Ratio (FLLR)

One good method to assess the behavior of the network feeder is through the power losses parameter. For a three-phase feeder, if the total load in a phase p is P_p and the total real power loss across each phase p of the feeder is $P_{loss}^{\ p}$, then the feeder loss-to-load ratio, FLLR for phase p can be expressed in Equation 4.

$$FLLR^p = \frac{P^p_{loss}}{P^p} = \frac{P^p_{loss}}{\sum_{k=1}^{n} P^{kp}} \qquad (4)$$

It is expected that with PV integration, power loss will be reduced as a result of less power flow from the substation feeder. However, under the opposite scenario of reverse power flow, the power loss across each phase of the feeder may increase.

4) Substation Reserve Capacity Index (SRCI)

From the loading level of a substation or distribution transformer, we can determine its remaining reserve capacity available. Thus, a network performance indicator can be developed in order to assess the reserve capacity on the substation transformer. To determine the loading level on the low-voltage side of the three-phase, the apparent power flowing through the segment connecting from the substation node to the first node of the feeder can be used. The ratio of the sum of individual apparent powers in each phases of the feeder, $S^a{}_{01}$, $S^b{}_{01}$ and $S^c{}_{01}$, to the capacity of the substation transformer, S^{sub}, provides the substation loading level, which can be used to calculate the substation reserve capacity index, SRCI, as given in Equation 5.

$$SRCI = 1 - \frac{|S^a_{01}| + |S^b_{01}| + |S^c_{01}|}{S^{sub}} \qquad (5)$$

With PV penetration levels on the LV network, the loading level of the substation transformer may decrease and

operate at a lower efficiency, leading to a larger value for SRCI. On the other hand, with extremely high PV inclusions, and under reverse power flow scenario, the value of SRCI can also increase [15].

B. Case Study: Mauritius

In this study, a portion of the LV distribution network of Mauritius was taken as a case study for the simulations. A detailed survey of the distribution network was carried out, involving the main distribution transformer, the main LV feeder type and length, and the distribution of consumer loads on the network as shown in Table I and II.

TABLE I. DISTRIBUTION TRANSFORMER DATA

Distribution Transformer Data	
Parameter	***Description***
Model	SIEMENS
Rating	250 kVA
Ratio	22 kV/415V

TABLE II. LV FEEDER DATA

LV Feeder Data	
Parameter	***Description***
Type	XLPE Torsade
Size	3 x 70 mm^2 + 1 x 54.6 mm^2
Current	213 A
Length	423 m

Consumption data was taken from a previous survey carried out by the electric utility, for typical medium-class residential consumer loads (Fig. 4). This data was then input as the typical profile for each load in the model.

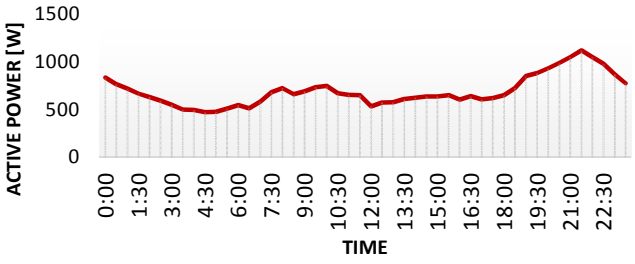


Fig. 4: Consumer Active Power Load Profile

The active load demand varies from a minimum of 473 W to a maximum of 1,119 W and consists of two peaks, a morning peak and an evening peak as shown in Fig. 4. The reactive power profile was derived from the active power load profile since this data is normally not recorded by energy meters and residential customers are billed only based on active power consumption by the utility.

C. Model configuration

The representative LV distribution network was modelled in DigSILENT PowerFactory Model (Version 2018) as shown in Fig. 5. The network comprises of one MV/LV distribution transformer and its downstream connections to main household nodes represented by the blue circles, to which, internally, one or more consumers are connected.

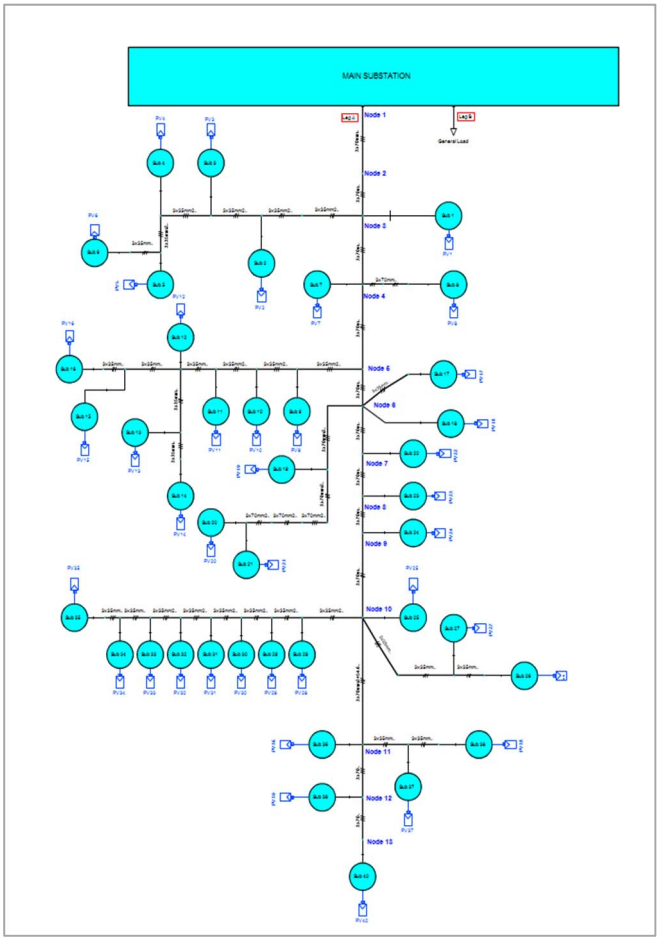


Fig. 5: Test Model of LV Distribution Network in DigSILENT

The grid supplies electricity to 136 residential consumers through a 250 kVA 22/0.415 kV Dyn11 distribution transformer. To every household node having a cluster of LV loads connected to the grid, a single-phase Photovoltaic system was connected to the LV terminal (230V) of a single-phase load. The interconnection of several of these PV systems to the same distribution LV feeder in a small area is normally defined as a PV 'cluster' [16]. Based upon existing PV systems installed on the Mauritian LV distribution network, the maximum PV capacity has been set as 2.5 kW per household node in the model.

The LV loads and cluster of PV systems were balanced as far as possible over the three phases a, b and c of the feeder distribution line using the DigSILENT advanced function '*Phase Balance Optimisation*'.

A load test (results shown in Table III) was carried out on the two outgoing LV feeders of the existing 250kVA distribution transformer at noon (time at which peak generation from PV systems occur) in order to calibrate and validate the developed model on DigSILENT against the real measured parameters.

TABLE III. LOAD RESULTS (SOURCE: LOAD TEST)

Transformer Load Test Results					
Name	St Helene	**Time**	12h00	**% LOAD**	45 %
Current	**Phase a**	**Phase b**	**Phase c**	**Neutral**	**Total kVA**
Feeder A	72	146	64	78	67
Feeder B	75	81	21	57	42
					109

D. Load Flow

For each voltage control mode of the PV inverter, a load flow was run on DigSILENT for different scenarios involving increasing penetration level of PV on the grid of up to 250%.

The only constant parameter involved in all load flows was the load demand, the value of which was fixed at noon, as this represents the time of peak PV generation due to maximum irradiance level. Additionally, another scenario involved running the load flow for a daylong variation of the load and analyzing the effect on the distribution network through the previously defined performance indicators.

1) Scenario 1: IEEE 1547 Mode

Under this mode, the PV is not allowed to participate in voltage regulation and does not contribute any reactive power to the grid. The grid PV penetration was increased in steps of 25% up to a maximum of 250% as mentioned above with a maximum PV capacity of 2.5 kW per household node for this scenario and the following scenarios below.

2) Scenario 2: Q(V) Characteristic Mode

Here, all PVs in the model are allowed to inject or absorb reactive power from the grid. However, the amount of injected/absorbed reactive power depends on the voltage at the terminal to which the PV is connected. This Q-V dependency is defined by the characteristic curve as shown in Fig. 6 and this operating mode characteristic was configured on the PVs in the model.

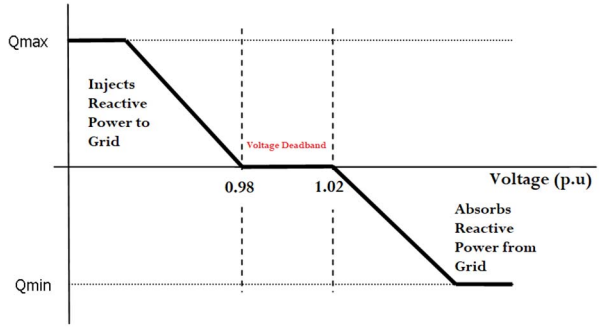


Fig. 6: Q-V Characteristic curve of PV inverter

Within the voltage deadband limit, the PV inverter does not inject nor absorb any reactive power. A value for the deadband in the range of 0.98 and 1.02 was chosen based on the desired sensitivity of reaction of the PV inverter. Once the terminal voltage goes outside of the deadband limit, the PV inverter will either inject or absorb reactive power as illustrated in Fig. 6. The slope outside the voltage deadband limit is referred to as the droop, chosen to be 4 %, in this simulation and allows a gradual injection or absorption of reactive power.

3) Scenario 3: Qmax Mode

Under this mode, the PV inverter is allowed to inject maximum reactive power to the grid irrespective of the voltage at the PCC.

V. RESULTS

The first simulation is to show how the voltage can drop along the feeder from the distribution transformer to the far end without any PV connected as shown in Fig. 7.

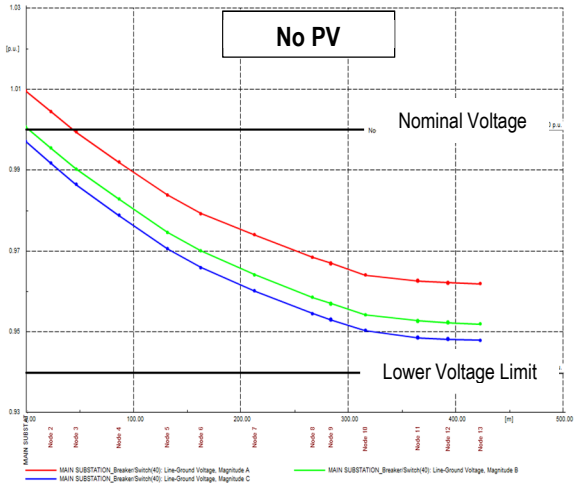


Fig.7: Voltage Profile along feeder – No PV

The voltage profile was taken from simulation results obtained after running a load flow on DigSILENT with the voltage for each phases of the feeder plotted on the graph with respect to the feeder length. The voltages for the three phases drop very close to the lower voltage limit allowed by the utility.

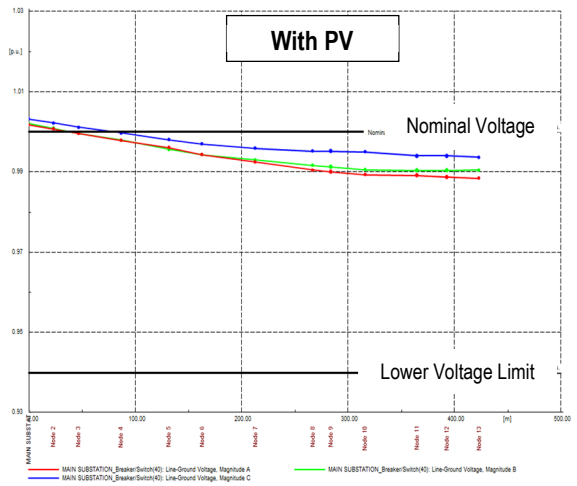


Fig. 8: Voltage Profile along feeder – With PV

With PV integration into the grid system, it can be seen that the voltage profile is improved and increased well above the lower voltage limit, even at the far end of the LV feeder as shown in Fig. 8.

A. Reactive Power Flow Variation Under Different Control Strategies

Since the voltage control modes would largely influence the reactive power flow in the grid, the study was carried out to determine how the reactive power flow on the grid changes in relation with different voltage control modes adopted by the PV inverter. The PV rating was oversized to add reactive power capability to the inverter, while providing maximum active power. The results obtained are shown for this simulation is shown in Fig. 9.

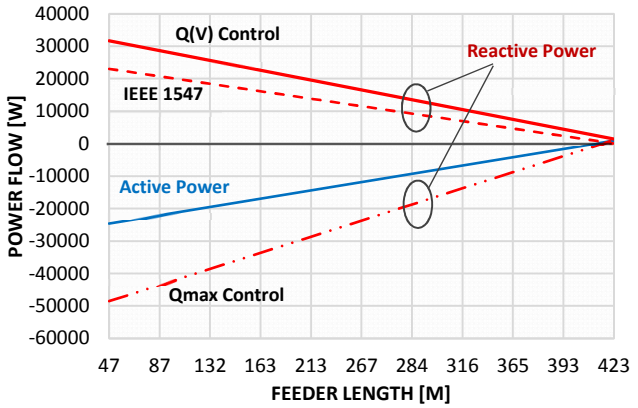


Fig. 9: Power Flow along feeder for different PV Voltage Control Modes

The flow of active and reactive power will obviously decrease as we go downstream of the feeder, but the slopes are different for each control mode of the PV inverters. Under the IEEE 1547 mode, the PV inverters injects only active power whereas reactive power is taken from the grid (red dotted line) as shown in Fig 10. With the Q(V) control approach, the PV inverters absorb reactive power to regulate the voltage at the terminal to which they are connected whenever the voltage deadband threshold value is exceeded. This lead to a greater flow of reactive power in the feeder (red solid line). With the Qmax approach, PV inverters deliver maximum reactive power and excess reactive power, after meeting the reactive load demand, flows upstream back to the network grid.

B. Network Performance Indices Variation with high PV Penetration

The penetration level of PV into the feeder was then varied from 0% (Without PV) to 250% (with PV capacity of 2.5 kW per household node) in incremental steps of 25% (steps of 250W). During the simulation, the load profile was fixed at noon where the active power demand is 533 W and the reactive power demand is 175 VAr and PV generation is maximum. The network performance indicators MVDI, AFLI, FLLR and SRCI were then evaluated against varying PV penetration levels and the results obtained are as shown in Fig. 10 (a), (b), (c), and (d), respectively.

For both the IEEE 1547 and Q(V) mode, the MVDI decreases linearly until, at larger penetration PV levels, the Q(V) mode stabilizes to near zero deviation from the nominal voltage due to voltage control action from the inverters as shown in the MVDI plot in Fig. 10(a). This is done at the expense of an increased reactive power flow as shown in the AFLI plot in Fig. 10(b). On the other hand, operating the inverters under the Qmax mode (green line) cause maximum voltage deviation.

With increasing PV penetrations in the grid, the average feeder loading flowing downstream tends to decrease linearly at first and then increases slightly at the instant at which reverse power flow occurs. In comparison to the other modes of operation, the average feeder loading increases more under the Qmax mode as shown in the AFLI plot in Fig. 10(b).

The feeder losses as shown in the FLLR plot in Fig. 10(c) also follow the same trend as the average feeder loading by decreasing at first and reaching nearly zero losses in the feeder at the instant of reverse power flow.(a)

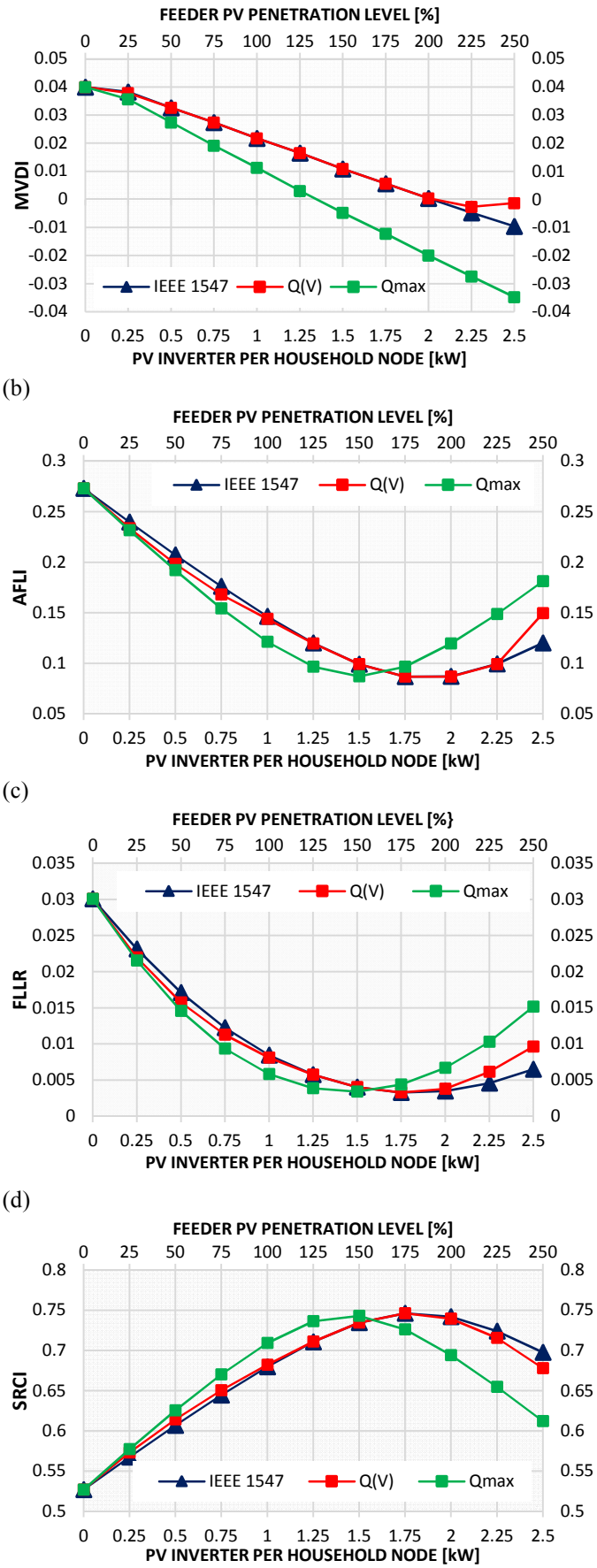
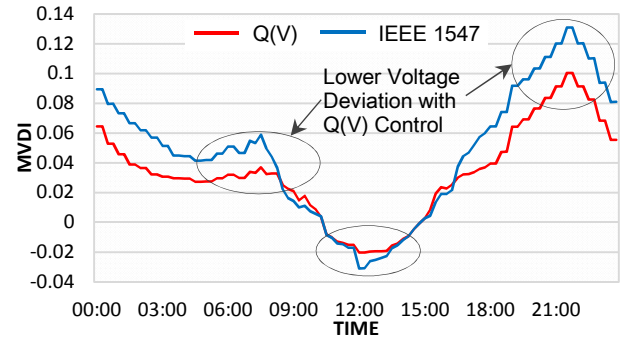


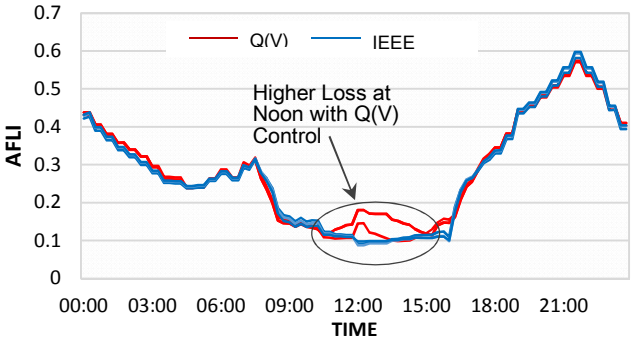
Fig.10 The four performance indices against sizing of PV inverter; a) MVDI, b) AFLI, c) FLLR, and d) SRCI

Due to increased feeder loading, the reserve capacity of the substation transformer also varies as shown in the SRCI plot in Fig. 10(d). The reserve capacity decreases after a threshold value of PV penetration, more steeply when operating in the Qmax mode than for the other modes. The threshold value of PV penetration at which the value of SRCI is maximum is 175% for both the IEEE 1547 (blue line) and Q(V) mode (red line) and 150% for the Qmax mode (green line).

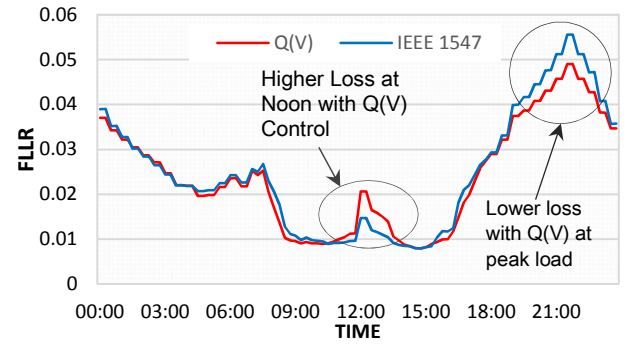
(a)



(b)



(c)



(d)

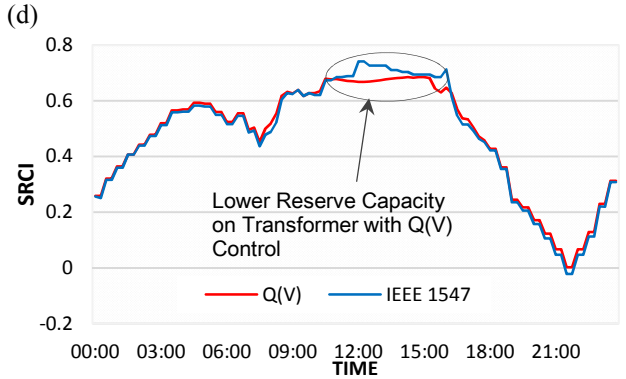


Fig. 11: Day-long variation: (a) MVDI; (b) AFLI; (c) FLLR; (d) SRCI

C. *Daylong Variations of Network Performance Indices*

The daylong variation of the network performance indices was analyzed by running a Quasi-Dynamic simulation in DigSILENT. The consumer load variation for a 24-hour period was applied with a 30-min resolution. The PV profile used was actual data, with a 15-min resolution, taken from an operational PV farm on the Mauritian grid and scaled down to match a maximum capacity of 2.5 kW per household. The following results of daylong variation for each performance index are obtained as shown in Fig. 11(a), (b), (c) and (d).

The MVDI plot in Fig. 11(a) shows that the maximum voltage deviation parameter is lower for the Q(V) control approach for times of the day when there is no PV penetration from midnight to 08h00 and also lower at the peak load demand in the evening. This value is also lower for the Q(V) control mode at the start of the day when the PV output is near its minimum value. Furthermore, for time of peak output of the PV at 12h00, the Q(V) control shows a somewhat better approach than the IEEE 1547 approach to regulate the voltage at the far end of the LV feeder.

As shown in Fig. 11(b) AFLI plot, the daylong variation is somewhat the same for both control approaches except at noon, where the Q(V) control increases the feeder loading for all three phases as compared to the IEEE 1547 approach.The FLLR Plot in Fig. 11(c) shows higher feeder losses for the Q(V) control approach as compared to the IEEE 1547 approach at the time of peak output generation of the PVs at noon due to increased feeder loading. On the other hand, at times of peak load demand, the Q(V) approach does a better job by decreasing the feeder losses.

As indicated in the SRCI Plot in Fig. 11(d), the reserve capacity on the substation transformer decreases at noon due to reverse power flow resulting from the PV generation. The Q(V) approach shows a much greater decrease on the reserve capacity as compared to the IEEE 1547 approach.

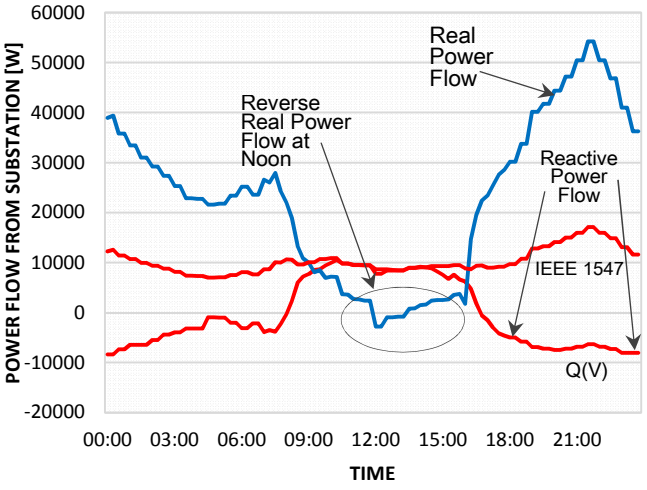


Fig. 12: Daylong Variation of Active and Reactive Power Flow

The variation of active power flow from the substation transformer is shown in Fig. 12 and as indicated above, the real power is reversed at noon, going beyond the negative zone, where PV generation is maximum. Active power is almost the same for both the IEEE 1547 approach and the Q(V) control approach. However, the reactive power flow is higher for the IEEE 1547 approach than for the Q(V) control approach almost at all times of the day.

VI. DISCUSSIONS

From the first part of the analysis, it is observed that for all control approaches, IEEE 1547, Q(V) and Qmax, the network performance indices AFLI, FLLR and SCRI tend to either increase or decrease after a threshold value of 1.5 – 1.75 kW equivalent to 150 – 175% of PV penetration. However, the MVDI value tends to decrease for all three control approaches as soon as the PV penetration level reaches 25% and reaches a zero value at 125% (Qmax) and 200% (IEEE 1547 and Q(V)) of PV penetration. These variations are summarized in Table IV.

TABLE IV. PERFORMANCE INDICES WITH PV PENETRATION

Performance Index	Performance Indices with PV Penetration		
	Control Approach	*Variation*	*PV Penetration (%)*
MVDI	IEEE 1547	Decrease	25 %
	Q(V)		25 %
	Qmax		25 %
AFLI	IEEE 1547	Increase	200 %
	Q(V)		200 %
	Qmax		150 %
FLLR	IEEE 1547	Increase	175 %
	Q(V)		175 %
	Qmax		150 %
SRCI	IEEE 1547	Decrease	175 %
	Q(V)		175 %
	Qmax		150 %

Considering all the above parameter variations, in order to minimize feeder losses and loading, operate the substation transformer more efficiently and have a voltage value near to the nominal value at the far end of the LV feeder, the optimal PV capacity for a household node, which can be integrated into the grid is 1.75 kW.

The second part of our analysis on daylong variation demonstrates that the network performance indices are affected most at noon, time at which PV output is maximum, which is in line with the results obtained in the first part of the analysis. Moreover, the Q(V) control mode helps to decrease the feeder losses much better than the IEEE 1547 approach at times where the load demand is highest (in the evening) and also regulates the voltage much better at times when there is no PV generation. In terms of reserve capacity on the substation transformer, Q(V) control impose a lower reserve capacity than IEEE 1547, thereby increasing the operating efficiency of the transformer. The main concern is that Q(V) control tends to increase the feeder losses and loading much more than IEEE 1547 at noon. As a matter of fact, a properly devised control logic/ algorithm adapted to the operation mode of PV inverter and using a combination of both the Q(V) and IEEE 1547 control approach, will ideally optimize the performance of the distribution network.

On analyzing the power flows from the substation transformer, it can be observed that the Q(V) control approach compensates for the flow of reactive power from the substation transformer to meet the reactive demand as compared to the IEEE 1547 approach.

VII. CONCLUSIONS

In this paper, the performance of the distribution network has been studied under high PV penetration. Four performance indicators were formulated to analyze the network performance in terms of voltage deviation, feeder loading, feeder losses and transformer reserve capacity. A portion of the distribution network of Mauritius was surveyed and modelled in DigSILENT software to carry out the simulations.

The analysis show that increasing the PV penetration level into the test grid up to a certain level can be favorable in terms of grid support, and if we go beyond the threshold value, the network performance indices tend to worsen due to reverse power flow. From the discussion section, this threshold value for the optimal PV capacity has been identified to be 1.75 kW per node to which consumers are connected.

Further simulations determined the day long effect on the network from the participation of the PV inverters to regulate the reactive power. The Q(V) control mode was found to be most effective to decrease reactive power flow from the substation transformer, feeder losses during evening peak load demand and maximize operating efficiency of the substation transformer.

ACKNOWLEDGMENT

The author gratefully acknowledges the contributions and support of the electric utility of Mauritius, the Central Electricity Board (CEB) in the making of this report. Furthermore, the author recognizes the DigSILENT support team, who have also provided assistance whenever needed to develop the model.

REFERENCES

[1] M. Schachinger, *Module Price Index*. [online] Available at: https://www.pv-magazine.com/features/investors/module-price-index/, Aug. 2018.

[2] Ministry of Renewable Energy & Public Utilities, Long-Term Energy Strategy 2009-2025, pp.3,21,35, Oct. 2009.

[3] *Home Solar Project,* [online] Available at: http://ceb.intnet.mu/resultsdraw/resultsdraw.asp, May 2018.

[4] E. Demirok, D. Sera, P. Rodriguez, and R. Teodorescu, "Enhanced local grid voltage support method for high penetration of distributed generators," IECON 2011 - *37th Annual Conference of the IEEE Industrial Electronics Society in Melbourne, Australia*, pp. 2481–2485, Nov. 2011.

[5] International Electrotechnical Commission, "Interconnecting distributed resources with electric power systems," *IEC/IEEE/PAS 63547*, ed1.0, Sep. 2011.

[6] P. Esslinger and R. Witzmann, "Evaluation of reactive power control concepts for PV inverters in low-voltage grids," *CIRED 2012 Workshop: Integration of Renewables into the Distribution Grid, Lisbon, Portugal*, May 2012.

[7] "IEEE Application Guide for IEEE Std 1547, IEEE Standard for Interconnecting Distributed Resources with Electric Power Systems," *IEEE Std 1547.2-2008,* pp. 13, Apr. 2009.

[8] K. Turitsyn, P. Sulc, S. Backhaus, and M. Chertkov, "Local Control of Reactive Power by Distributed Photovoltaic Generators," *2010 First IEEE International Conference on Smart Grid Communications in Gaithersburg, MD, USA*, pp. 79–84, Jun. 2010.

[9] T. Stetz, M.Rekinger, and I.T. Theologitis, "Transition from Uni-Directional to Bi-Directional Distribution Grids," *Photovoltaic Power Systems Programme*, pp. 8-9, Sep. 2014.

[10] E. Liu, J. Bebic, G. G. Research, and N. R. E. Laboratory, Distribution system voltage performance analysis for high-penetration photovoltaics: National Renewable Energy Laboratory, p. vii, Feb. 2008.

[11] J. Gómez-González, D. Cañadillas-Ramallo, B. González-Díaz, J. Méndez-Pérez, J. Rodríguez, J. Sánchez, and R. Guerrero-Lemus, "Reactive power management in photovoltaic installations connected to low-voltage grids to avoid active power curtailment," *Renewable Energy and Power Quality Journal*, vol. 1, pp. 7–8, Apr. 2018.

[12] P. Carvalho, P. Correia, and L. Ferreira, "Distributed Reactive Power Generation Control for Voltage Rise Mitigation in Distribution Networks," *IEEE Transactions on Power Systems*, vol. 23, no. 2, pp. 766–772, May 2008.

[13] Y. Liu, J. Bebic, B. Kroposki, J. D. Bedout, and W. Ren, "Distribution System Voltage Performance Analysis for High-Penetration PV," *2008 IEEE Energy 2030 Conference in Atlanta, Georgia USA*, p. 3, Nov. 2008.

[14] R. Walling, R. Saint, R. Dugan, J. Burke, and L. Kojovic, "Summary of Distributed Resources Impact on Power Delivery Systems," *IEEE Transactions on Power Delivery*, vol. 23, no. 3, p. 1, Jul. 2008.

[15] M. Alam, K. M.. Muttaqi , D. Sutanto, L. Elder & A. Baitch, "Performance analysis of distribution networks under high penetration of solar PV," *44th International Conference on Large High Voltage Electric Systems in France 2012*, pp. 3-6, 2012.

[16] Y. Ueda, K. Kurokawa, T. Itou, K. Kitamura, Y. Miyamoto, M. Yokota, and H. Sugihara, "Performance Ratio and Yield Analysis of Grid Connected Clustered PV Systems in Japan," in *Photovoltaic Energy Conversion, Conference Record of the 2006 IEEE 4th World Conference in Waikoloa, HI, USA*, pp. 2296-2299, May 2006.

Cloud Robotic Architectures: Directions for Future Research from a Comparative Analysis

Viraj Dawarka
Faculty of IT, Design and Communication
Curtin Mauritius, Charles Telfair Campus
Telfair, Moka, Mauritius
vdawarka@curtinmauritius.ac.mu

Girish Bekaroo
School of Science and Technology
Middlesex University Mauritius
Coastal Road, Uniciti, Flic-en-Flac, Mauritius
g.bekaroo@mdx.ac.mu

***Abstract*— Advances in robotics and cloud computing have led to the emergence of cloud robotics where robots can benefit from remote processing, greater memory and computational power, and massive data storage. The integration of robotics and cloud computing has often been regarded as a complex aspect due to the various components involved in such systems. In order to address this issue, different studies have attempted to create cloud robotic architectures to simplify representation into different blocks or components. However, limited study has been undertaken to critically review and compare these architectures. As such, this paper investigates and performs a comparative analysis of existing cloud robotic architectures in order to identify key limitations and recommend on the future of cloud robotic architectures. As part of this study, 7 such architectures have been reviewed and compared and results showed limited evaluation of existing architectures in favour of security weaknesses.**

Keywords*— *Cloud-Robotic Architectures; Robotics; Cloud Computing; Comparative Analysis; RaaS;

I. INTRODUCTION

With the expeditiously increased capacity of artificial intelligence methods, improvements in microprocessors and development in cloud platforms, robot technology have climbed up the ladder in terms of potential and automation. Fundamentally grounded on cloud and robotic technologies, the emergence of cloud robotics enables the fusion between infrastructure cloud empowered by machine-to-cloud (M2C) communications and an ad-hoc cloud formed by machine-to-machine (M2M) communications among cooperative robots [1]. In other words, cloud robotics is the evolution of conventional robotics technology such that when connected to the cloud, facilities such as robust computational power, limitless data storage and communication resources are obtained from the state of the art data center found in the cloud that can process and share data from diverse robots and agents [2]. During recent years, the conceptualization of cloud robotic architectures has provided means for automation in large scale systems. This has been facilitated by automating robots with sensory and actuation capabilities that can capture data publish to the cloud for post processing [2]. The use of appropriate architectures in addition to cloud systems have also shown to provide a myriad of benefits [3]. Firstly, access to remote libraries for robots is made easily available through big data. Also, through collective robot learning, trajectories could be easily shared between robots and cloud computing gives access to on-demand statistical analysis.

Due to these benefits, several cloud robotic architectures have been developed during the past decade. The creation of such architectures also provide a simplistic representation of the complex underlying structures of cloud robotics. Some architectures are made up of components such as Service-Oriented Architecture (SOA) which uses web services to communicate between the cloud and the robots whereas some use client and cloud side techniques and even the incorporation of Internet of things in the architecture [4]. In terms of research related to cloud robotic architectures, few published literature is available that establishes the comparison between a ranges of implemented architectures. As related work, a previous study was followed on the cloud robotics architectures, its associated challenges and applications [1] but provided limited comparative analysis between existing architectures. Another study provided survey of related work in the field of cloud robotics and automation [4] with over 75 references in the area. Nevertheless, the study reports five ways how cloud robotics and automation can improve performance. Within this study, improvement of performance was addressed in the following areas: Big data, Cloud computing, collective robot learning, open source and open access, and crowdsourcing and call centers but the major scope was not on the comparative analysis of existing cloud robotic architectures. To address this gap, this research paper investigates and performs a comparative analysis of existing cloud robotic architectures in order to identify key limitations and recommends on the future of cloud robotic architectures.

This paper is structured in the following consecutive way: The first section gives an overview and introduction of the topic, then the second section describes the methodology used to achieve the purpose of this paper, followed by a review of current cloud robotic architectures given in the third section. In the fourth section, a comparative analysis of the existing cloud robotic architectures is provided and in fifth section provides the recommendations on future cloud robotic architectures, before concluding the paper in the sixth section.

II. METHODOLOGY

The purpose of this research paper was accomplished by adapting methodologies used in previous studies related to comparative analysis of architectures [1, 2]. The process started by a comprehensive paper search involving different online databases including Google, Google Scholar and online research

databases (IEEE, ACM, Springer and Elsevier). These platforms were utilized because of their popularity and the plethora of papers published. Ultimately, Google and Google Scholar were used to target any other papers not published within the mentioned research databases. The search process involved different keywords including mainly "cloud robotics" and "cloud robotics architecture". From an initial selection of 68 conference and journal articles to be reviewed thoroughly, 7 cloud robotics architectures were identified and critically analyzed. Once the architectures were selected, literature search for each particular architecture was thoroughly conducted through further investigation using relevant articles and related key websites. The information obtained was then analyzed thoroughly to write-up this paper and are presented in the next sections. This methodology was utilized in different studies conducting comparative analysis [1, 2].

III. Review of Cloud Robotic Architectures

Based on the methodology defined in the previous section, different cloud-robotic architectures were identified and are further discussed as follows:

A. Cloud Robotics Architecture by Terrissa and Ayad

A previous study [5] proposed a new cloud robotics architecture where robots can be provided as a service easily, efficiently and cheaply. The purpose of this architecture is to provide a way for users of robots (e.g. household, military robots, etc.) to utilize on-demand cloud platforms as a runtime environment of their operating systems, while also permitting to customize tasks of robots without interaction with the provider. The architecture is composed of two parts, namely, the client side and the cloud side. The client side consists of modules that occur at the client including a client administrator interface and communication module that links the mentioned interface and robots. The cloud side has modules involved in the Cloud robotics provider platform including a cloud robotics administrator interface, Virtualization Layer, Virtual Robot Systems, among others. Compared to classic cloud computing architectures, this proposed architecture has an integrated Virtual Robot Layer which consists of two components, namely a Robot Management System and a Virtual Robot System. The former is designed to manage and control robots from web whilst the latter is the operating system of the robot that is executed virtually onto the cloud environment. Within the architecture, three types of actors are present, namely the Client Administrator, Cloud Robotics Administrator and Cloud Administrator. The client Administrator takes the responsibility of the management and configuration of the local robots via a web-based interface. On the other hand, the Cloud Robotics Administrator operates mainly on the Virtual Robotics layer and finally, the cloud administrator manages the cloud infrastructure.

B. Cloud Robotics by Wan et al

In the study by Wan et al [3], cloud robotics architecture is viewed as an arrangement of two parts, namely the cloud infrastructure and the bottom facility. Whilst the cloud platform consists of key equipment including servers and database as shown in Fig. 1, the bottom facility includes robotic equipment in the form of mobile robots and unmanned aerial vehicles, among others. In terms of key features, the cloud infrastructure provides various benefits such as dynamic computing tasks and elasticity of resources. Furthermore, most of the processing is conducted in the cloud, whilst being facilitated by networking devices. Moreover, computational load could be shifted to the cloud for processing thus resulting into smaller robot loads which also benefit from longer battery life.

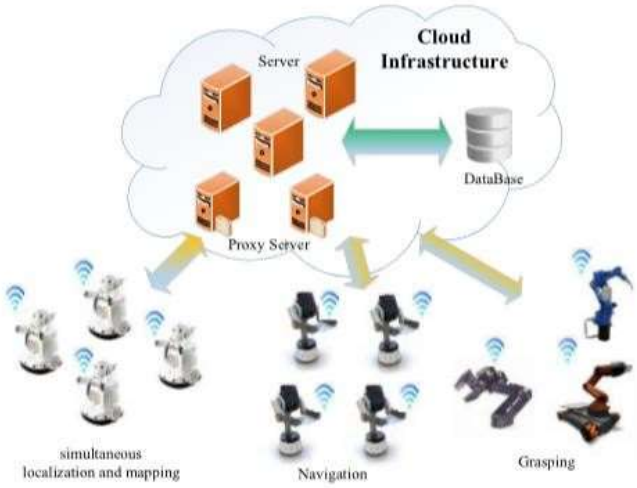


Fig. 1 - Cloud Robotic Architecture by Wan et al [11]

C. Internet of Things (IoT) Infrastructure-as-a Service (IaaS) Architecture by Mouradian et al

In order for robots to virtualize robots and to allow them to upstream applications as a service, Mouradian et al proposed an Internet of Things (IoT) Infrastructure-as-a Service (IaaS) [4]. This architecture was conceptualized so as to extend another existing architecture [5] while providing benefits such as flexibility, elasticity and cost efficiency. The architecture consists of four layers as shown in Fig. 2, namely, Network-Level Virtualization Layer, Node-level Virtualization Layer, Physical Robots Layer and Gateway Layer. The top-most layer, namely the Network-Level Virtualization Layer interfaces with the IaaS platform and consists of different modules such as a Publication Engine and Robot Monitor, among others. The Node-level Virtualization Layer is made up of different virtualized robots and the Physical Resources Layer consists of the supported robots. Finally, the Gateway Layer aims to conceal the heterogeneity and specificities of the robots such as involved communication protocols, APIs, among others. The proposed architecture was also implemented in the earthquake search and rescue domain to implement a fire suppression feature by robots. For evaluation, the prototype was compared with a peer-to-peer overlay network and results showed that the proposed architecture meets a set of key requirements.

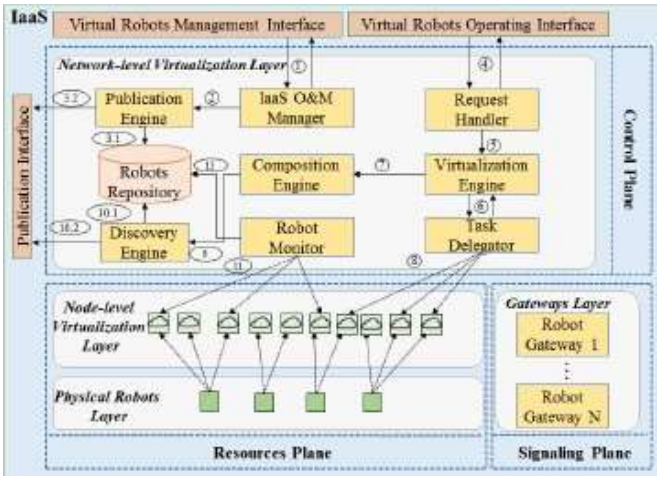


Fig. 2 - IoT IaaS Architecture by Mouradian et al [4]

D. The RoboEarth Systems Architecture

Announced in 2009, the RoboEarth project envisioned "a World Wide Web for robots: a giant network and database repository where robots can share information and learn from each other about their behavior and environment" [3]. The architecture consists of three components namely the Server, Generic Components and Robotic Specific. The Server deals with database-related objects including images, the environment (maps and actions) and also web services for reasoning. The Generic components is built-up of four parts. Firstly, action and situation recognition and labelling that facilitate the generation of action recipes. Secondly, Action Execution ensures that action is executed on the robots through proper coordination of the RoboEarth database. Thirdly, Environment Modelling combines existing information from RoboEarth database and robots sensor and fourthly, the Semantic Mapping which uses Simultaneous localization and mapping (SLAM) to mix observations of environment with identified objects from the database. Finally, a learning module is present to allow knowledge to be obtained in the form of feedback based on the performance of robot's work. The third component, Robot Specific, consists of hardware abstraction layer which allows interaction between the computer and the robot through means of drivers and motion primitives.

E. Cloud-Based Robot Grasping System Architecture by Bekris et al.

Bekris et al. [4] presented a system architecture for Cloud-Based object recognition and grasping, consisting of two phases, namely, the offline and the online. The offline phase consists of three sections, specifically Cloud, Humans and Robots/Humans. In the cloud component, there are the Google Object Recognition Engine, the Google Cloud Storage which sends Computer Aided Design (CAD) models for analysis to the Grasp Analysis section which in term sends the Candidate Grasps to the Google Cloud Storage. The Human component has two sections, that is, Label and Domain Knowledge. The Label interacts with the Google Object Recognition Engine to train the images with the object labels whereas the Domain Knowledge sends all the semantic data such as CAD Model, Center of Mass, weight, texture and material to the Google Cloud Storage. The last components, Robots/Humans deal with the camera where image training occurs with the Label. In the offline system, the images of each object are stored so as it can be trained with the object recognition server. The object is then utilized for the creation of a grasp which is further analyzed to know the robustness to spatial uncertainty. The second phase is the online phase which consists of two components namely, the Cloud and Robots. The cloud section contains the Google Object Recognition Engine connected to the Google Cloud Storage through the Object Label. On other hand, the Robot component has the camera which sends images to the Google Object Recognition Engine, in addition to the 3D Sensor which points cloud to the Pose Estimation. The final module is the Select Feasible Grasp with Highest Success Probability which sends the grasp execution results to the cloud storage and in terms received the candidate grasps. The execution of the online phase allows the robots to analyse the results and store them onto the cloud server for further references.

F. Integrated Service-Oriented Architecture with Robot as a Service by Chen et al

In their architecture, Chen et al [10] defined the concept of Robot as a Service (RaaS), grounded on the Service Oriented Architecture (SOA). The architecture consists of three major blocks, particularly the RaaS unit, RaaS in the cloud environment and the interfacing devices to SOA. The key blocks of the architecture include the RaaS unit and the RaaS cloud. The RaaS unit is a service broker, where clients can search for services and applications accessible within the directory of the unit. The RaaS cloud consists of different applications deployed to the units. To implement the concept, a prototype was implemented, which functions using the basis that each unit or robot hosts a repository of preloaded services. Units can benefit from the RaaS cloud which contains applications deployed by developers or clients. To enable communication between the RaaS units and services in the cloud, different interfaces were implemented. Experiments were conducted to evaluate the prototype and architecture and results highlighted effective software and hardware system to support the complex underlying infrastructure.

G. IoT-based Cloud Robotic Architecture by Karnouskos et al

In a recent work [6], Karnouskos et al demonstrated integrated an Internet of Things (IoT) layer within a proposed cloud robotic architecture. The architecture was motivating such that IoT devices could be endowed with open source software in order to perform automated tasks by robots. The architecture consists of three major components as shown in Fig. 3, namely, the Cloud layer, Robot Operating System (ROS) nodes, and the Things layer. The cloud layer contains key modules including Data Lake, IoT service and Web Dashboard through which an end-user can interact. ROS nodes contains a series of independent processes namely Workflow Engine, ROScore, Thing Integrator and Thing Controller. The final layer relates to the "things" of IoT and could be different Internet-enabled objects. In order to implement the architecture, a prototype was also created so as to show autonomous robot interactions and behavior while also enabling enterprise integration.

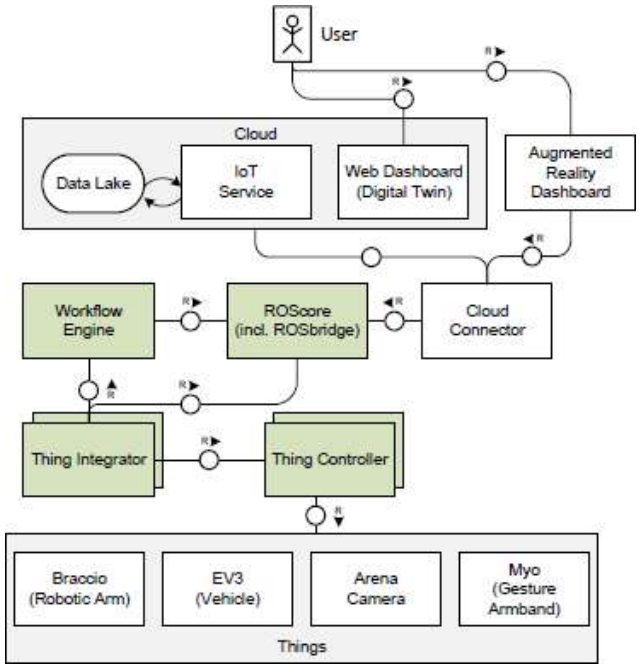


Fig. 3 - IoT-based Cloud Robotic Architecture [6]

IV. COMPARATIVE ANALYSIS

The review of cloud-robotic architectures shows that most of them consist of four components, namely, a cloud administrator, a client, cloud and a network. The key feature among the architectures studied relate to the use of cloud platform to store data and makes access of information between the robots easier, which aligns with the basis of cloud robotics. All the architectures were designed with a different purpose to address some focused problem as listed in **Error! Reference source not found.**. For instance the architecture by Mouradian et al focused on IoT and IaaS, whilst the one proposed by Karnouskos et al focused on the integration of OSS technologies with IoT within the architecture. The IoT-based Cloud Robotic Architecture by Karnouskos et al was also found to be significantly different as compared to others as it provides virtualization of robots which other architectures do not use as a service. When the same architecture is compared against the Cloud-Based Robot Grasping System Architecture by Bekris et al, the latter uses real robots for obtaining data such as images for grasp analysis. Having the new trends of facilities incorporated within the architecture leads to a major disadvantage which can discourage people of using as there is dependence between nodes where if one node fails, the other process cannot run smoothly.

The Cloud-Based Robot Grasping System Architecture by Bekris et al makes a difference when compared against existing architectures through the use of grasping which others do not use. Despite having common attributes such as Cloud for storage purposes and the use of robots to gather data, this architecture enables object recognition through the use of Google Object Recognition Engine. The grasp analysis which is performed by this architecture allows to determine the robustness to spatial uncertainty and pose estimation to select a grasp for reference. On the other hand, the architecture by Terrissa and Ayad and the one by Chen et al have the Service-Oriented Architecture as basis of the proposed architecture. As claimed in these studies, the use of such underlying architecture improves communication with the cloud through the use of web services. The only difference between the two architectures is that the one by Terrissa and Ayad has not been tested onto a real cloud infrastructure. On the other hand, the architecture proposed by Chen et al has an increase overhead due to service interaction with another service as a SOAP and REST protocol is being used. It has also been observed that SOA is not desirable when developing real-time services as asynchronous communication occurs between the services. Contrarily, the RoboEarth Systems Architecture was found to have a massive advantage compared to the other architectures as it is equipped with an immense database where large amount of data are stored and robots can retrieve data for analysis. Paradoxically, having this main advantage, it lacks privacy and security concerns when it concerns cloud connectivity. Thus, it can discourage this architecture to be used if the database is handling confidential data. A comparative summary of the architectures analysed are given in TABLE I.

TABLE I. COMPARATIVE SUMMARY

Cloud Robotic Architecture	Purpose	Type of Architecture	Key Components	Key Features	Key Limitations
A. Cloud Robotics Architecture by Terrissa and Ayad	• Allows robots to conduct several work through connection of clouds and not relying on the features or hardware of robots.	Service-Oriented Architecture	• SOA(Service Oriented Architecture) Paradigm • ROS (Robot Operating system) • Cloud Administrator • Client • Cloud • Network	• The SOA paradigm enables the use of web services to communicate with the cloud. • ROS is the operating system of the robots • Cloud Administrator deploys new services on the cloud to a client. • Client can be either the administrator who configures the robots or the robots who which benefits the services. • Cloud stores all the packages and acts as the infrastructure • Network allows the communication between the cloud and the client.	• Poor Network connection can lead to delay in communication • Based on SOA paradigm and ROS as a robotic middleware thus not tested onto real cloud infrastructure.

B. Cloud Robotics by Wan et al	• Allows multi-robot cooperative works using SLAM and navigation though the use of cloud infrastructure	Network architecture	• Cloud Infrastructure • Bottom facility	• Cloud Infrastructure: Composed of high performance of servers, large databases, proxy servers and other components. • Bottom facility: Consists of mobile robots, machinery, unmanned aerial vehicles and others.	• Large delay when using network robotics system
C. Internet of Things (IoT) Infrastructure-as-a Service (IaaS) Architecture by Mouradian et al	• It proposes an IoT IaaS architecture which enables virtualization of robots and provides them as-a-service for applications • It enables the use of robots for flexibility, elasticity and cost-efficiency making into consideration the advantages of cloud such as scalability and virtualization.	IaaS (Infrastructur e as a Service) and Virtualization architecture	• PaaS & SaaS Domain • Robots Service Marketplace • IaaS Domain • Gateway Domain • Physical Robots Domain	• In the PaaS & SaaS Domain, a Google app engine is used to host and execute the application. • Robots Service Marketplace domain provides the presence server. • IaaS Domain provides four RESTful web services using Java Restlet framework for two robot services. • In the Gateway Domain, a robot gateway is present to map the IaaS HTTP java REST with the robots API. • Physical Robots Domain consists of the robots namely the LEGO Mindstorms NXT which are being used in architecture.	• Difficulty in homogenizing the same node-level virtualization procedures for the entire IoT resource. • Existing SLA (Service Level Agreement) and QoS management procedures in the IaaS not able to handle robots services due to mobility.
D. The RoboEarth Systems Architecture	• It comprises of a massive database and network where the robots can share details and can learn from one another regarding their environment and behavior.	Robotic and Embedded system	• Server • Generic Components • Robot Specific	• The server consists of several types of databases namely the Object database, Environment database, Web services and Action Databases. • Generic Components have the object recognition, action and situation recognition and labelling, Action execution, master control, Data encoding, environment modelling and Learning. • Robot Specific consists of the hardware abstraction layer.	• Privacy and security concerns regarding the cloud connectivity. • Potential of robots being attacked remotely.
E. Cloud-Based Robot Grasping System Architecture by Bekris et al.	• It enables the estimation of robustness to spacial uncertainty through use of robots and cloud using grasping.	Online and Offline system of Cloud & Robot grasping architecture	• Cloud (Google Object Recognition Engine, Google Cloud Storage, Grasp Analysis) • Robots (Camera, Label, Domain Knowledge, 3D Sensor, Pose Estimation)	• The system architecture consists of Offline phase and online phase. • The offline phase functions with the recording of digital photos onto an object recognition server. A 3D CAD model of every object is issued and a candidate grasp is generated where each grasp is analysed to estimate the robustness to special uncertainty. • The online phase works as when a photo is taken from the robot and sent to the object recognition server through network. The server then issue an object for the stored data. A 3D point set is sued to measure the robot for pose estimation and a grasp is selected from a pool of grasps available. After analysis, the robot stores the result onto cloud for reference.	• 3D Point sets when doing pose estimation. • A better analysis of image recognition is needed as it is difficult when dealing with false positive and false negative. • Low confidence values with images. • Grasp analysis issues when dealing CAD Models.
F. Integrated Service-Oriented Architecture with Robot as a Service by Chen et al	• It proposes a robot to be an all-in-one SOA unit and can communicate with the cloud.	Service-Oriented Architecture	• A service Provider, A Service Broker, A Service Client • Generic hardware based Intel architecture, USB and Serial Port	• The RaaS unit consists of services and applications Directory. • The software in the RaaS communicate with drivers, hardware and Operating system.	• Increase Overhead due to service interaction with another service. • SOA is not desirable for Real-time as services communicate asynchronously.

			• Graphic composition based on Robotics Developer Studio and VPL Language	• The RaaS units can communicate with each other through Wi-Fi or Bluetooth. • The RaaS unit communicates with other services in the cloud through standard service interface WSDL. • Communication protocol used is SOAP or REST Protocol.	
G. IoT-based Cloud Robotic Architecture by Karnouskos et al	• It brings forward the aspects of autonomous behavior, decentralization, object detection, decision-making, track & trace integration with enterprise system. • It allows the developers familiarize with new OSS technologies pertaining IoT integration.	Cloud and IoT Architecture	• Cloud • Augmented Reality Dashboard • ROSnodes • Cloud connector • Things	• Cloud consists of a data lake linked with the IoT service and a Web Dashboard. • The Augmented Reality Dashboard visualizes data via an iPad pointed to devices or designated locations. • ROSnodes consists of a collection of independent process such as Workflow Engine, ROScore, Thing Integrator and Thing Controller. • Cloud connector use the ROSbridge to get access to data available via ROScore. • Things consist of a Braccio (Robotic Arm), EV3(Vehicle), Arena camera and Myo(Gesture Armband)	• The architecture relies on each components to function, if one node fails, the flow will not work.

V. THE FUTURE OF CLOUD ROBOTIC ARCHITECURES

Through the research conducted with multiple types of cloud robotic architectures implemented so far, it was observed that more work needs to be done to critically evaluate each architecture, while also comparing and benchmarking between them. The different papers reviewed in this study revealed that most architectures have not been experimentally validated against many factors that affect the real world. This could be due to limited availability of evaluation frameworks meant for cloud robotic architectures. Also, it was found that architectures based on SOA paradigm and ROS as a robotic middleware have not been tested onto real cloud infrastructure thus not making the architecture viable and efficient. A new architecture also needs to be implemented where each component does not need to rely on each other making the architecture independent to function and having a good flow with the communication from the robots to cloud.

Some studies also reported a large delay when using network robotic system and this was found to slow-down communication between the cloud and robots. This problem could be addressed by reducing the length of messages interchanged between the robots and the cloud. Additionally, limited architectures have integrated IoT components and more focus could be put in the IoT IaaS architecture in regards to mobility. This is because using this architecture causes existing Service Level Agreement difficult to handle thus resulting to loss within the system using it.

In this era where security and privacy about data is regarded as a key aspect, the architecture such as the RoboEarth needs more security concerning cloud connectivity so as the data being communicated is secure and not leaked. Furthermore, the operation over the Internet and networks also imply that such systems inherit the vulnerabilities of operating in these environments. As such, more security testing of these architectures is needed. In addition, the computer security research community need to propose cloud robotic security architectures and standards to improve security of robots operating in these modes against different types of attacks.

VI. CONCLUSIONS

This paper analyzed and compared 7 cloud robotic architectures proposed by researchers and roboticists. From the review, it was found that the Service Oriented Architecture has been commonly used as basis of some cloud robotic architectures due to improved communication as well as the ease of integration with web services. Moreover, recent architecture has also been designed to accommodate the Internet of Thing aspect where the communication flows from cloud to robots through use of nodes in between. However, different gaps were identified during the review where the major one was that limited evaluation were conducted for these architectures, while also testing large scale applications. In addition, more work is needed to ensure low delay network traffic communication from cloud to robot and even an efficient real-time processing architecture. Overall, although some cloud robotics architectures have been proposed, various avenues for future work remain available to address the identified and future insights given in this paper. As future work, implementation of the architectures could be considered for practical analysis, comparison and benchmarking.

REFERENCES

[1] G. Hu, W. Tay and Y. Wen, "Cloud robotics: architecture, challenges and applications," IEEE network, vol. 26, no. 3, pp. 21-28, 2012.

[2] B. Krishna, J. Oviya, S. Gowri and M. Varshini, "Cloud robotics in industry using Raspberry Pi," in IEEE Second International Conference on Science Technology Engineering and Management (ICONSTEM), Chennai, India, March 2016, pp. 543-547

[3] S. Miratabzadeh, N. Gallardo, N. Gamez, K. Haradi, A. Puthussery, P. Rad and M. Jamshidi, “Cloud robotics: A software architecture: For heterogeneous large-scale autonomous robots,” in World Automation Congress (WAC), Rio Grande, Puerto Rico, 4 October2016.

[4] B. Kehoe, “Cloud-based methods and architectures for robot grasping,” UC Berkeley, Berkeley, California, 2014, Available at: ProQuest ID: Kehoe_berkeley_0028E_14814. Merritt ID: ark: /13030/m56d8xmn. Retrieved from https://escholarship.org/uc/item/898085r2

[5] C. Mouradian, S. Yangui and R. Glitho, “Robots as-a-service in cloud computing: search and rescue in large-scale disasters case study,” 15th IEEE Annual Consumer Communications & Networking Conference (CCNC), Las Vegas, Nevada, USA, 12-15 January 2018.

[6] A. Tang, J. Han and P. Chen, “A comparative analysis of architecture frameworks,” in 11th Asia-Pacific Software Engineering Conference, Busan, Korea , 03 December 2004, pp. 640-647

[7] T. Magoulas, A. Hadzic, T. Saarikko and K. Pessi, “Alignment in enterprise architecture: A comparative analysis of four architectural approaches,” Electronic Journal of Information Systems Evaluation, vol. 15, no. 1, pp. 88, 2012.

[8] B. Kim and R. Neff, “Measurement and communication of greenhouse gas emissions from US food consumption via carbon calculators,” Ecological Economics, vol. 69, no. 1, pp. 186-196,2009.

[9] A. Zakari, A. Lawan and G. Bekaroo, “Towards Improving the Security of Low-Interaction Honeypots: Insights from a Comparative Analysis,” in International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering, Mauritius, 25-27 November 2016, pp. 314-321

[10] L. Terrissa and S. Ayad, “Towards a new cloud robotics approach,” in 10th IEEE International Symposium on Mechatronics and its Applications (ISMA), Sharjah, United Arab Emirates, 8-10 December 2015, pp. 1-5

[11] J. Wan, S. Tang, H. Yan, D. Li, S. Wang and A. Vasilakos, “Cloud robotics: Current status and open issues,” IEEE Access, vol. 4, pp. 2797-2807, 2016.

[12] B. Kehoe, S. Patil, P. Abbeel and K. Goldberg, “A survey of research on cloud robotics and automation,” IEEE Transactions on Automation Science and Engineering, vol. 12, no. 2, pp. 398-409, 2015.

[13] K. Bekris, R. Shome, A. Krontiris and A. Dobson, “Cloud automation: Precomputing roadmaps for flexible manipulation,” IEEE Robotics & Automation Magazine, vol. 22, no. 2, pp. 41-50, 2015.

[14] Y. Chen, Z. Du and M. García-Acosta, “Robot as a service in cloud computing,” in Fifth IEEE International Symposium on Service Oriented System Engineering (SOSE), Jiangsu, China, 04-05 June 2010, pp.151-158

[15] S. Karnouskos et al., "Experiences in integrating Internet of Things and cloud services with the robot operating system," 2017 IEEE 15th International Conference on Industrial Informatics (INDIN), Emden, 2017, pp. 1084-1089.

Simultaneous Localization and Mapping in Application to Autonomous Robot

Olusanya Agunbiade and Tranos Zuva
Faculty of Applied and Computer Science
Vaal University of Technology
South Africa
agunbiadeoy@tut.ac.za, tranosz@vut.ac.za

***Abstract*—The important characteristic that could assist in autonomous navigation is the ability of a mobile robot to concurrently construct a map for an unknown environment and localize itself within the same environment. This computational problem is known as Simultaneous Localization and Mapping (SLAM). In literature, researchers have studied this approach extensively and have proposed a lot of improvement towards it. More so, we are experiencing a steady transition of this technology to industries. However, there are still setbacks limiting the full acceptance of this technology even though the research has been conducted over the last 30 years. Thus, to determine the problems facing SLAM, this paper conducted a review on various foundation and recent SLAM algorithms. This study was carried out to discuss unresolved SLAM problem with the view that could encourage researchers to produce more innovative achievements by taking the SLAM issues discussed into consideration. However, a novel SLAM technique that will address these problems will be proposed.**

Keyword—autonomous robot, illumination variance, dynamic environment, navigation, simultaneous localization, mapping

I. INTRODUCTION

Simultaneous Localization and mapping (SLAM) is an important problem that has been broadly researched in robotics. Its contribution towards autonomous robot navigation has attracted researchers towards focusing on this area [1, 2]. In the past, various techniques for addressing simultaneous localization and mapping has been proposed with remarkable achievements. Research has been conducted based on several types of sensors because of their advantage to one another; also sensor selection can be as a result of the type of technique they are proposing to solve the SLAM problem [3]. In [4], they proposed to address the SLAM problem in an environment without pre-determined landmarks using laser sensors.

The Particle filter algorithm was deployed on their system and had produced a highly detailed map for an office environment. But, laser sensors apart from being expensive could produce a wrong measurement when encountering shiny or black objects that do not reflect light and this could affect robot localization in an environment [5]. In [2], they proposed the use of the Extended Kalman Filter (EKF) filter algorithm to process the information obtained by the sonar sensors attached to the robot. The sonar sensor was proposed because of its low cost and low computational complexity for retrieving information from the environment. In the work of [3], they condemn the use of sonar sensors because of its inability to provide fine-grained information from sound. Instead, they proposed bio-sonar. This was employed due to high intelligent interaction capability towards a complex environment and its ability to extract more information from the echoes than sonar. BatSLAM algorithm was proposed to analyse the information acquired by the bio-sonar, but from their experiment, the system limitation occurs if it encounters a larger complex environment because echoes arriving from different directions are delayed and trying to analyse them produced an invalid cue which makes the system fail to navigate correctly [6].

In the work of [7], they proposed a vision-based SLAM, because the camera was able to acquire more information from the environment than other sensors which could improve robot navigation [8]. Thus, being aware of the issues of environmental noise with vision-based system, they tend to address the issue of shadow. They proposed the use of two-dimensional occupancy grid maps produced from 3-D point clouds obtained by a stereo camera, they also introduced an extracted salient line segments from the ground into the grid map. On the grid map, robot pose estimation was attained by employing particle filters. In this technique, the grid maps were not affected by shadow and lighting conditions, but under severe illumination condition, it is impossible to extract the salient line segment which resulted to a failed SLAM technique. However, the issue of illumination variance support while some researchers still prefer the use of active sensors to acquire data from the environment [9].

Furthermore, for reliable and accurate measurement of the environment, the work of [10] propose the use of multiple sensors to attain an impressive result so that one sensor can take advantage over the weakness of another. But the major limitation is high computational complexity when combining too much data from multiple sensors. Hence, irrespective of the sensor employed, they all have their limitations. However, sensors are not the only contributor to SLAM failure, algorithm employed to address the SLAM problem also have their limitations and they will be discussed in the next section. In this paper, Section 2 and 3 discuss the foundation and recent SLAM algorithm while section 4 discusses the proposed Novel SLAM technique.

II. FOUNDATIONAL SLAM ALGORITHM

In an attempt to develop an efficient and effective technique that can address SLAM problem, sensors play an important role of acquiring data from the environment [8] but localization and mapping techniques are not limited to this operation. There are several procedure that still need to be implemented, for instance, the analysis of the data captured assist in mapping building and localization, this can be attained by using SLAM algorithm [11]. In the literature, several foundation SLAM algorithms has been proposed with outstanding result but they are all confronted with various challenges and issues [12]. In this section, some of these algorithms will be discussed together with their limitations and advantages.

2.1 Extended Kalman Filter (EKF)

In the review involving EKF, it is important to mention the Kalman filter because it is the foundation for EKF and some algorithms like Extended Information Filter (EIF), non-linear least-square etc. [12]. Researchers over the years have employed Kalman filter as an algorithm to estimate dynamic linear systems with Gaussian noise [11]. Kalman filter represent a state vector (μ_t) given in equation (1).

$$\mu_t = (S_t, l_1, l_2 ...., l_N) \quad (1)$$

It is formulated by estimating the landmark (l) where N represents numbers of map landmark, the current pose (S_t), as well as a covariance matrix $(\sum_t)$ which signifies the covariance between the state. However, the issue of non-linearity in the robot model is the limitation of the Kalman filter algorithm and trying to address this issue led to Extended Kalman filter [12].

Extended Kalman Filter (EKF) is an upgraded version of a Kalman filter that can address non-linear model [11]. The linearization of non-linear model can be solved by many methods but in EKF, a technique called first order Taylor expansion is employed to address this issue. At each time (t), it linearize the measurement and motion model using the current state to estimate for a new update [11]. The filter procedure is attained with two steps given in section 2.1.1 and 2.1.2.

2.1.1 The Time Update stage

At this stage, the filter computes the covariance matrix $\hat{\sum}_t$ and the predicated state $\hat{\mu}_t$ at time (t). Expressions is given in equations (2) and (3).

$$\hat{\mu}_t = f(\mu_{t-1}, u_t) \quad (2)$$

$$\hat{\sum}_t = A_t \sum_{t-1} A_t^T + G \wedge_u G^T \quad (3)$$

where A_t signifies the Jacobian of the motion model f as related to the robot pose S_t that is evaluated at μ_t, u_t represent the robot control, $\wedge_u$ signifies the covariance matrix related to this stage and G represent a projection matrix.

2.1.2 The measurement update stage

This stage plays a significant role to address the problem of data association (c) and generate the new updated measurement for μ_t and $\sum_t$ using the current state of the previous stage. They are computed by estimating first, the Kalman gains as given in equations (4)

$$K_t = \hat{\sum}_t C_t^T \left(C_t \hat{\sum}_t C_t^T + \wedge_Z \right)^{-1} \quad (4)$$

$$\mu_t = \hat{\mu}_t + K_t \left(z_t - h\left(\hat{\mu}_t, c \right) \right) \quad (5)$$

$$\sum_t = (I - K_t C_t) \hat{\sum}_t \quad (6)$$

Where I represent identity matrix, K_t represents the Kalman gain, C_t represent the Jacobean of the measurement model h in relation to every detected land marks estimated at $\hat{\mu}$ and the pose of the robot, $\wedge_z$ signifies the covariance matrix as related to this stage and z_t represent the sensor measurement [11].

The above listed two steps are unique to EKF and can be use to instantiate the online SLAM given a condition of Gaussian model. The EKF is a popular algorithm because of the ability to overcome the problem of Kalman filter, but the computational cost of the algorithm is high [13].

2.2 Rao-Blackwellized Particle Filter

In the work of Murphy, Rao-Blackwellized particle filter is implemented to estimate the joint posterior $(p(S_{1:t}, m | Z_{1:t}, U_{1:t-1}))$ of a map (m) and the robot trajectory $(S_{1:t} = S_1, ..., S_t)$. The estimation is carried out given the odometry measurement $(U_{1:t-1} = U_1, ..., U_{t-1})$ and the observation measurement $(Z_{1:t} = Z_1, ..., Z_t)$ of the robot. Using the information provided above, the Rao-Blackweilized particle filter used the factorization given in equation (7) to represent SLAM [14].

$$p(S_{1:t}, m | Z_{1:t}, U_{1:t-1}) = p(m | S_{1:t}, Z_{1:t}) . p(S_{1:t} | Z_{1:t}, U_{1:t-1}) \quad (7)$$

In the factorization procedure, firstly the trajectory of the robot is estimated followed by the map taking into consideration that the same trajectory is computed. This is mandatory because the map creation rely on the estimated robot pose and vice versa [22]. The Rao-Blackwelized particle filter offers efficient computational cost with improved processing speed. Its representation in equation (7) can be formulated efficiently to address SLAM since the posterior of the map $p(m | S_{1:t}, Z_{1:t})$ will be estimated by employing the technique of mapping with known poses given that $S_{1:t}$ and $Z_{1:t}$ are known [14].

In computing, the posterior $p(S_{1:t} | Z_{1:t}, U_{1:t-1})$ for potential trajectories, particle filter may be employed. Since every particle corresponds to a potential robot trajectory and individual maps are connected with each sample. Therefore, maps will be created using the observation and corresponding particle relating to the trajectory. This procedure allows the robot to learn models of their environment and estimate successfully their trajectory [14]. Thus, the particle filters effectiveness and complexity rely heavily on the number of particles. The increase in the number of particles might improve its effectiveness, but at a price of high computational cost. Otherwise, the effectiveness can be minimized with low computational cost. However, estimating an optimal number of particle required is often difficult to attain [15].

In the literature, the above mentioned foundation SLAM algorithms are very common [9] but not limited to these three

algorithms, others such as, Unscented Kalman filter (UKF) and Compressed Extended Kalman Filter (CEKF) exist [12]. The algorithms discussed above are the basis of operation for other algorithms [11]. Irrespective of the algorithm proposed, they all have their advantages and limitations, and these shows that more research still need to be done in this area. However, researchers do not only rely on these algorithms, they tend to introduce more components that can improve SLAM performance, some of their methods will be presented in the next section

III. RECENT SLAM ALGORITHMS

The simultaneous localization and mapping is an important problem to consider as far as autonomous guidance is concerned [1]. Thus, successful navigation of a robot given in an unknown environment require continuous updating of map while simultaneously estimating its position in its environment. In literature, researchers have presented recent technique to overcome the challenges of the foundational SLAM algorithm [16]. This section will discuss some of these techniques, their challenges and issue they encountered during their implementation. Furthermore, this section will disclose recent issue because reviews are only conducted on recent techniques only.

In the work of [17], they propose to replan at every time (k) when there is a probability distribution update on the state of the autonomous robot, this technique is referred to as: Simultaneous localization and Planning. The algorithm employed to carry out this task is known as Partially Observable Markov Decision Process (POMDP). This algorithm was proposed because of its ability to cope with uncertainties and changes. The idea employed in motion planning under uncertainty is to identify a policy (π_k) at each time (k) that generate control state (U_k) using the available information of the robot. Thus, there are other important terms that needs to be defined to achieve this goal. Given a robot in an unknown environment whose state is represented by x_k at time step (k), the motion noise and control state at time (k) is represented by W_k and U_k respectively. The state evolution model can now be formulated using equation (8)

$$x_{k+1} = f(x_x, U_k, W_k) \tag{8}$$

However, in any partial observable system, the sensor vector measurement at very time k represented as Z_k plays an important role of providing observation measurement. The expression for Z_k is given in equation (9).

$$Z_k = h(x_k, V_k) \tag{9}$$

where V_k denotes sensing noise.

On this note, the data available for deriving decision at each time k is the history of controls and observation as expressed in equation (10), the conditional probability distribution for the overall possible robot state is given in equation (11)

$$\mathrm{H}_k = (Z_{0:K}, U_{0:k-1}) = (Z_0, Z_1, \ldots\ldots Z_k, U_0, \ldots\ldots, U_{k-1}) \tag{10}$$

$$b_k = p(x_k | H_k) \tag{11}$$

The b_k generated in equation (11) is also referred to as information state or belief that compressed the data H_k and can be recursively computed using the last state and current observation as expressed in equation (12)

$$b_{k+1} = \alpha p(Z_{k+1} | x_{k+1}) \int_X p(x_{k+1} | x_k, U_k) b_k dx_k \tag{12}$$

where α represent the normalization constant and the U_k can be generated based on the information state using a policy π_k as expressed in equation (13)

$$U_k = \pi_k(b_k) \tag{13}$$

The π_k represent the solution of a POMDP over a continuous observation space with a limitation that is intractable. In addressing this issue, Feedback based Information Road Map (FIRM) was proposed to minimize the intractable problem to a tractable POMDP by generating a representative graph in the information state space.

Given a FIRM graph with controller, policy π^g can be extracted by mapping graph nodes (v) to the edge (m) as expressed in equation (14)

$$\pi^g : v \rightarrow m \tag{14}$$

Thus, the set of all graph planar (Π^g) generated in an information state space allows POMDP to be tractable on the FIRM graph as expressed in equation (15)

$$\pi^{g^*} = \arg\min_{\Pi_g} \mathrm{E} \sum_{n=o}^{\infty} C^g(B_n, \pi^g(B_n)) \tag{15}$$

where visited n-th node is represented by B_n., $C^g(B_n, \pi^g(B_n))$ signifies the cost function. Experimental performance attained is impressive, but the technique is unable to cope with dynamic environment. In their future work, they want to propose a frame work that can learn and model changes using prior knowledge of object motion.

In the work of [18] vision-based technique using a monocular camera was proposed to initiate SLAM. The algorithm employed in their technique is known as Parallel, Tracking and Mapping (PTAM). In PTAM, procedures are split into two level task operations in parallel. In the tracking level, monocular camera fixed on mobile robot is used to capture images from the environment. Ground feature-based pose estimation algorithm was proposed to detect ground features. But to achieve a more accurate robot pose, a weighted projection error-based energy function expressed in (16) was used to achieve this task.

$$\min_{\mathrm{x}} \sum w(r_p) \| r_p(\mathrm{x}) \| \tag{16}$$

where w represent the tukey biweight function for the homography-based projection error (r_p).

Given the accurate robot localization attained by the expression in (16), matched features triangulated was used to generate an initial map. Afterwards is the second level task known as mapping thread, during this task, the initialized map is queried for incorporating new key features and these is achieved by employing the use of epipolar searching procedure. The new matched features are selected for acceptance by searching for candidate region around the epipolar with minimal differences of Zero-Mean Summed

Squared Differences (ZMSSD), while candidate region with higher differences in ZMSSD compared to the threshold will be rejected for re-mapping. Expression is given in equation (17)

$$s\left(u^{i}_{j}\right)=\begin{cases} r_j \le \delta_i & accept \\ r_j \rangle \ \delta_i & reject \end{cases} \quad (17)$$

where

$r_j = \pi\left(H_i v^{i}_{j}\right) - u^{i}_{j}$,

H_i represent the homography estimated by Random Sample Consensus (RANSAC) algorithm, v^{i}_{j} signifies the reference feature of u^{i}_{j}, r_j represent the projection error and δ_i represent the threshold.

After classification using equation (17), new map points generated are in-cooperated into the map points to improve the accuracy of the system. Indoor experimental performance carried out shows tremendous achievement towards accuracy, but from the future work presented in their study, they intend to improve the performance of their technique to cope with mapping in various illumination scenarios.

The work presented in the study of [19] is a graph based simultaneous localization and mapping. The concept of graph-based SLAM relies on representing the nodes present in the graph by each pose attained by the robot. In real world scenario, these nodes can be used to signify features extracted from images captured by camera sensors or laser point cloud. Nodes can also be employed to signify physical landmark of object like trees, cars etc. Edges present in the graph is signified by a factor connecting two nodes. These factors represent the bearing measurement of features. Given a robot navigating in an unknown environment using graph-based SLAM, the first problem to address is creating a graph. This can be attained by identifying the nodes and the factors connecting these nodes based on the data generated from the sensors, this computation is known as front-end. The second problem to address is the nodes configuration that provides best explanation for the factors. These steps assist to compute a maximum likelihood map and these computations is known as back end. Thus, the back ends aim to find configuration of nodes that minimize error created by the factors from the front-end operation. If $x = \left(x_1, ..., x_n\right)^T$ represent a state vector and x_i signifies pose of node i which can also represent a robot or landmark position. The error function $\left(e_{ij}\left(x\right)\right)$ description for a single factor between i and j represent the different between Z_{ij} and $\hat{Z}\left(x_i, x_j\right)$. Expression is given in equation (18).

$$e_{ij}\left(x\right) = \hat{Z}\left(x_i, x_j\right) - Z_{ij} \quad (18)$$

where $\hat{Z}\left(x_i, x_j\right)$ signifies the expected measurement given in the current state, Z_{ij} represent the observed measurement, i and j represent the graph nodes. The re-projection error of the observed landmark must be minimized for accuracy purpose. Thus, the minimization error can be expressed in equation (19)

$$\mathrm{x}^{*} = \arg\min_{x} \sum_{ij} e_{ij}\left(x\right)^T \sum\nolimits_{ij}^{-1} e_{ij}\left(x\right) \quad (19)$$

where $\sum_{ij}^{-1}$ represent information matrix related to the factors that exist between two pose x_i and x_j, $\sum_{ij}$ signifies the covariance matrix and x^{*} represent the optimal configuration of nodes with limited error induced from the factors of front end operation. However, the effectiveness of graph based technique towards SLAM has attracted researcher like [19]. In their enhanced graph-based SLAM, experimental result shows tremendous success towards SLAM problem, but at an expense of high computational cost due to an increase in computational requirement at the matrix factorization stage. Furthermore, their graph-based SLAM couldn't cope with dynamic environment and in their future work, they will be improving the SLAM technique towards tracking of dynamic object.

Irie et al. proposed to address the problem of SLAM for outdoor navigation taking into consideration drastic illumination changes which happen in most environments [7]. In this technique, stereo camera capable of obtaining 3-D range data was employed to capture data from the environment. Afterwards, 2-D grid map that is not much affected by illumination condition is generated. Given the 2-D grid map, occupancy information and salient line segment can be extracted perfectly. The particle filter is employed to extract the robot pose while edge pint based stereo SLAM was used to obtain robot ego motion and the occupancy information simultaneously. This extracted information is used to address the SLAM problem. There are other important procedures carried out to develop their technique, the model in Figure 3 provides full description of the proposed SLAM technique for mobile robot navigation in outdoor environment.

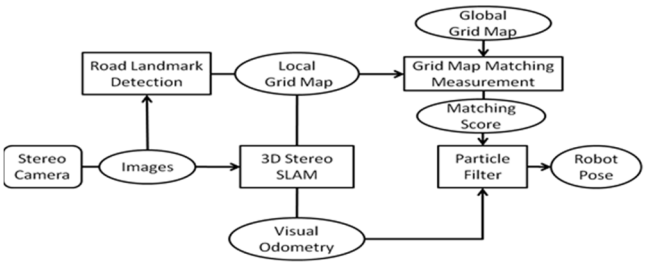


Figure 3: represent the stereo SLAM technique for mobile robot in outdoor environment [7]

The proposed model was successfully implemented and experimental performance of their visual odometry recovered from error and performed very well under various illumination situations. However, the technique failed under extremely adverse illumination condition such as when direct sunlight covers large part of the image and giving this condition, limited edge point is extracted for detection which resulted to huge error in motion estimation and inability to recover from kidnap robot.

IV. FUTURE WORK

Over the years, researchers have proposed various SLAM technique, but they were confronted with one problem or the other. In this review, some of the problems discussed are as follows: Illumination variation, kidnap robot, dynamic environment and high computational cost. Thus, all these problems if not fully addressed can be further minimize towards improving the SLAM performance. In the future work, filters could be employed to minimize the effect of illumination variation of shadow and light intensity. In

literature Normalize Difference Index is a common algorithm towards minimizing the effect of shadow [20] while dark channel prior and OTSU thresholding algorithms are used to minimize light intensity [5]. These filters could be operating in parallel since we are taking into consideration the issue of high computational cost. The kidnap problem which contribute to localization failure can also be addressed by introducing into the proposed SLAM the scan to match algorithm to search for the reference map of the current observation for re-localization of the robot [21]. The dynamic issue can be minimized by introducing the enhanced fuzzy clustering technique for keeping track of multiple dynamic object in the environment [22]. The SLAM algorithm is most likely to be particle based, because of it satisfactory result and acceptable performance towards computational cost. However, with the future SLAM technique possessing all these characteristics, it might offer better performance than any of the technique discussed in this review.

V. CONCLUSION

In this study, review on various SLAM algorithms was conducted, to understand how SLAM problem are being addressed, and the trend of problem encountered by researchers. SLAM has attracted many researchers because it supports the possibilities of concurrent execution of mapping and localization process. These become a great accomplishment in solving the problem of mobile robot autonomously achieving its goal without being controlled by anyone. More so, improving on this research area will require addressing problems associated with current SLAM. Therefore, review was conducted on recent and foundation SLAM to assist us discover persistence and recent problem associated to current SLAM techniques. However, the observation from this study and our previous review in [23] suggested that new researchers must focus their attention on the issue of high computational cost as related to processing time, this is a major problem mostly complained in SLAM. In addition, other problems such as illumination variance (light intensity and shadow), kidnap robot and dynamic environment are persistence problem mentioned by the researchers in their studies. Figure 5 shows the overall impact of SLAM problems as related to this research and in [23].

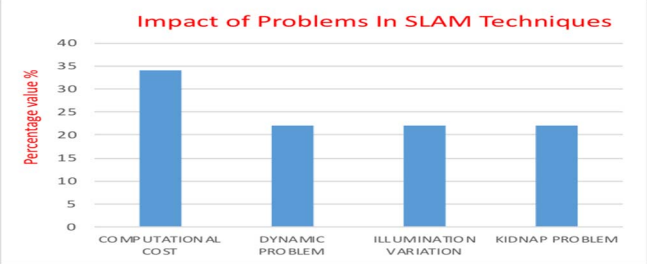


Figure 5: The overall impact of the SLAM problems

Furthermore, the advantages and limitations of sensors as related to SLAM problem were also discussed. In our future work, we will be implementing the proposed SLAM technique and successful execution will improve the performance of autonomous robot in path planning, exploration and mission planning.

REFERENCES

[1] J. Fuentes-Pacheco J, Ruiz-Ascencio, Rend, J. M., & N-Mancha, "Visual simultaneous localization and mapping: a survey," *Artif. Intell. Rev.* , vol. 43., 2015.

[2] G. Zunino and H. I. Christensen, "Simultaneous localization and mapping in domestic environments," *Multisensor Fusion and Integration for Intelligent Systems. MFI 2001. International Conference on,* pp. pp. 67-72., 2001.

[3] J. Steckel and H. Peremans, "BatSLAM: Simultaneous Localization and Mapping Using Biomimetic Sonar," *Plos One.,* vol. vol 8., pp. 1-11, 2013.

[4] A. Eliazar and R. Parr, "DP-SLAM: fast, robust simultaneous localization and mapping without predetermined landmarks," Proc. of the 18th International Joint Conference on Artificial intelligence. Acapulco, Mexico: Morgan Kaufmann Publishers Inc., 2003.

[5] O. Y. Agunbiade, S. M. Ngwira, and T. Zuva, "Enhancement optimization of drivable terrain detection system for autonomous robots in light intensity scenario," *The International Journal of Advanced Manufacturing Technology,* vol. 74, pp. 629-636, 2014.

[6] S. Hiryu, M. E. Bates, J. A. Simmons, and H. Riquimaroux, "FM echolocating bats shift frequencies to avoid broadcast–echo ambiguity in clutter.," *Proceedings of the National Academy of Sciences.,* vol. vol 107, pp. 7048-7053., 2010.

[7] K. Irie, T. Yoshida, and M. Tomono, "Outdoor Localization Using Stereo Vision Under Various Illumination Conditions," *Advanced Robotics,* vol. 26, pp. 327-348, 2012.

[8] S. Zehang, G. Bebis, and R. Miller, "On-road vehicle detection: a review," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 28, pp. 694-711, 2006.

[9] N. M. Thamrin, R. A. Adnan, R. Sam, and N. Razak, "Simultaneous Localization and Mapping Based real-Time inter-row Tree Tracking Technique for Unmanned Aerial Vehicle," 2012.

[10] J. A. Castellanos, J. Neira, and J. D. Tardos, "Multisensor fusion for simultaneous localization and map building. ," *IEEE Transactions on Robotics and Automation.,* vol. 17, pp. pp. 908-914, 2001.

[11] Y. Chen, "Algorithm for Simultaneous Localization and Mapping," pp. 1-15, February 3, 2013.

[12] S. E. Hadji, S. Kazi, and T. H. Hing, "A Review: Simultaneous Localization and Mapping Algorithms for Mobile Robot," *1st International Conference of Recent in Informtion and Communication Technologies,* pp. 83-92, 12th-14th September 2014.

[13] G. Dissanayake, S. Huang, Z. Wang, and R. Ranasinghe, "A review of recent developments in Simultaneous Localization and Mapping," in *2011 6th International Conference on Industrial and Information Systems*, 2011, pp. 477-482.

[14] K. M. Wurm, C. Stachniss, G. Grisetti, and W. Burgard, "Improved Simultaneous Localization and Mapping using a Dual Representation of the Environment," *Department of Computer Science, University of Freiburg, Germany,* pp. 1-10, 2003.

[15] M. Montemerlo and S. Thrun, "Simultaneous localization and mapping with unknown data association using FastSLAM," in *2003 IEEE International Conference on Robotics and Automation (Cat. No.03CH37422)*, 2003, pp. 1985-1991 vol.2.

[16] A. R. Khairuddin, M. S. Talib, and H. Haron, "Review on simultaneous localization and mapping (SLAM)," in *2015 IEEE International Conf. on Control System, Computing and Engg. (ICCSCE)*, pp. 85-90, 2015.

[17] A.-a. Agha-mohammadi, S. Agarwal, S. Chakravorty, and N. M. Amato, "Simultaneous Localization and Planning for Physical Mobile Robots via Enabling Dynamic Replanning in Belief Space," *Robotics: Systems and Control,* vol. abs/1510.07380, pp. 1-21, / 2015.

[18] S. Jia, K. Wang, and X. Li, "Mobile Robot Simultaneous Localization and Mapping Based on a Monocular Camera," *J. Robot.,* vol. 2016.

[19] P. Agarwal and W. Burgard, "Robust graph-based localization and mapping, PhD thesis,," *University of Freiburg,* pp. 1-155, 2015.

[20] O. Agunbiade, T. Zuva, A. Johnson, and K. Zuva, "Enhancement Performance of Road Recognition System of Autonomous Robots in Shadow," *Signal & Image Processing: An Inter. Jour.,* vol 4(6), 2013.

[21] S. Se, D. Lowe, and J. Little, "Vision-based mobile robot localization and mapping using scale-invariant features," in *Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No.01CH37164)*, 2001, pp. 2051-2058, vol. 2.

[22] A. Walcott-Bryant, M. Kaess, H. Johannsson, and J. J. Leonard, "Dynamic pose graph SLAM: Long-term mapping in low dynamic environments," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 1871-1878

[23] O. Agunbiade and T. Zuva: A Review: Simultaneous Localization and Mapping, doi:10.20944/preprints201805.0293.v1, 22 May 2018.

The Experimental Comparison of Features for Hand Detection

Frans Timbane, Shengzhi Du and Ron Aylward
Department of Electrical Engineering
Tshwane University of Technology
Pretoria, South Africa
fransthabo1@gmail.com

***Abstract*—Hand detection is critical in gesture recognition for conveying information or control commands between persons and computers. The accuracy of hand detection from images plays an important role in these applications. Extraction of effective features is the main factor in this task. The features should be discriminative, robust to different variations and easy to compute. This paper presents the experimental comparison of features commonly used in object detection, such as Haar-like features, a histogram of oriented gradient (HOG), and local binary pattern (LBP), using hand detection as the test platform. The adaptive boost (AdaBoost) cascade classification method is employed to combine "weak learners" to a strong classifier. The classifier was trained using 300 positive images, which are images containing the hand (region of interest (ROI)) and 10 000 negative images, which are images that do not contain a hand on them. Different parameter combinations of the classifier are considered for comparative experiments. The performance of the classifier using Haar, HOG and LBP features were evaluated with 320 static test images. The results show that parameter combinations have significant effects on the hand detection accuracy, which also differ when different features are used.**

Keywords—Hand detection, Haar-Like Features, Histogram of Oriented Gradient (HOG), Local Binary Pattern (LBP), Cascade classification.

I. INTRODUCTION

Object detection is a field of computer vision that deals with the detection of objects that belong to a certain class in a video or static image [1]. The object class work on the fact that every object has its own special features that are unique to that object which helps to identify that object. The ability to detect an object has a major impact on the application such as smart vehicles [2], video surveillance [3], security [4]. Object detection and recognition is one of the active research topics in image processing and computer vision field of study. Object detection is the process of detecting an object using the classification of data based on knowledge gained on features extracted from that object during training [5]. Feature extraction is one of the most important steps in the image description. Feature extraction converts pixels data into a representation of shape, motion, color, texture and spatial configuration of the object or its components [6]. Every feature extraction technique has its advantages and disadvantages. In this paper, we compare the performance of an adaptive boost (AdaBoost) cascade classifier on feature extractors such as Haar-like features, Histogram of Oriented Gradient (HOG) and Local Binary Pattern (LBP) through hand detection.

Haar Like features in object detection [7] was first introduced by Paul Viola and Michael Jones in Viola & Jones (2001). This method then attracted much attention from the researchers and it was extended to the various field of object detection. Haar-like feature extraction algorithms consider adjacent rectangular regions at a specific location in a detection window, sums up the pixel intensities in each region and calculate the difference between those sums [8]. HOG features were introduced by Dalal and Triggs. HOG extracts the shape information from the image, it divides the image into cells and calculates the occurrence of gradient orientation over them [9]. The HOG algorithm divides an image into small connected regions known as cells. For each cell a histogram of an oriented gradient is created by compiling the gradients of each pixel in a cell and using these gradients to vote in an edge orientation bin [10]. LBP operator labels pixels of the input image by thresholding neighbouring pixels with the centre pixel in a circle [11]. The LBP algorithm generates a binary code that describes the local texture pattern by normalizing the intensity values in the neighbourhood [12].

This paper is arranged as follows: Section II outlines the relevant work, section III illustrate the Viola and Jones Framework, section IV describes the methodology used for training the classifier, section V presents the experimental results and analysis, finally section VI presents the conclusion.

II. RELEVANT WORK

In many works of literature Haar-like features, the Histogram of Oriented Gradient (HOG) and Local Binary Pattern (LBP) are the most used feature extractors for cascade object detectors on digital images. These three approaches provide good results, real-time performance, and robustness. In most cases these approaches are commonly used for human feature (face) detection [13],[14],[15]. These methods can be used for the detection of any object. Other methods include SIFT [16], SURF [17]. The comparison study of feature extraction in face recognition system was conducted by N. Adinarayana, B. Ajanta Reddy and G. Balanagi Reddy using Haar based features. The comparison was based on extracting face features on a digital image under problems such as illumination, age, and surgery using Haar feature extractor. The results indicated that: under illumination on the first picture 530 features were extracted and on the second picture 548 features were extracted. On an aged image, 559 features were extracted on the first image and 622 features were extracted on the second image. Lastly, in a surgery image, 285 features were extracted on the first image whereas 283 features were extracted on the second image [18]. The results indicated that Haar extracted feature best on the aged picture.

Vivek Joy, Binu A, and Kuttyamma A.J presented the analysis for vehicle number plate detection using HOG and Hough Transform feature. The morphological operations are

used for the detection of the number plate. It came out that HOG performed better than Haugh Transform [19]. Dang Hoang, Sriprasertsuk Pao, and Kameyama Wataru conducted a study for fish detection and classification mechanism using LBP feature. The methodology used images which were captured by a non-stationary camera. Their proposed method achieves higher accuracy compared to the standard implementation, and it requires only 48% of the training time compared to the standard implementation [20]. Sorath Asnani, Sayed Danial Waseem, Ali Asghar Manjotho conducted a comparative study of Haar and HOG features using the same training parameters for detecting an unconcealed gun. The results show that Haar based feature took 105.6 hours which equals to 4.4 days while HOG based feature took only 2.5 hours. The performance indicated that HOG based classifier achieved the accuracy of 88.57% and the precision of 95.30%, whereas Haar based features achieved the accuracy of 42.14% and the precision of 45.73% [21]. on an unconcealed gun detection, HOG based features performed better than Haar based features.

Rushikesh Laxmikant Kulkarni conducted a research for handwritten character recognition using HOG, COM by Open CV and Python. HOG feature extractor has performed well for recognition of the characters, in different size and style and alignment with different background [22]. Cruz, Shinguemori, and Guimaraes have applied the comparison of HOG+SVM, Haar, and LBP to concrete and runway detection in high-resolution imagery [23]. The results indicated that Haar and LBP achieved the accuracy of 80%, however, the precisions were different, with LPB getting 78% and 50% for Haar. The HOG+SVM had a lower result, where the accuracy achieved 51% and a precision of 54%. D.K. Shedge, Blessy Mathai, Prasad Barude, Shivam Katkamwar, and Anjali Ranjan have designed a road sign detector using cascade classifier. They have used HOG as a feature detector since it can capture the shape of a road sign. The method of color segmentation was used in order to get the robust output. The experimental results indicated the detection method to be accurate. In the Image segmentation, they also achieved to remove the unnecessary background due to the fact that the color space has been used [24].

III. ADABOOST CASCADE CLASSIFIER FRAMEWORK

A. Cascade Classifier

Viola and Jones employed the AdaBoost method to combine a series of classifiers as a filter chain as shown in figure 1, which classifies image regions. The cascade eliminates false candidates if they don't pass the first stage. If they pass, then they will go to the next stage which is more difficult to pass than the previous stage. If the candidate passes all the stages then, it will mean that the object is detected [25].

B. AdaBoost

AdaBoost means Adaptive Boost, which refers to the method of training a boosted classifier [26]. AdaBoost consist of weak learners combined to form a boosted classifier in the form of equation (1).

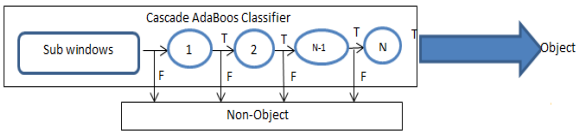


Fig. 1. Cascade Classifier

$$F_T(x) = \sum_{t=1}^{T} f_t(\mathrm{x}) \qquad (1)$$

where each f_τ is a weak learner that take x as the input and return a real value that indicates a class of an object. Each weak learner produces an output, hypothesis $h(x_i)$ for each sample in a training set. At each iteration, a weak learner is selected and assigned to a coefficient $\propto_t$ such that the sum of the training error at each stage is minimized.

$$E(t) = \sum_i^{\infty} E[F_{t-1}(x_i) + \propto_t h(x_i)] \qquad (2)$$

where $F_{t-1}(x_i)$ is the boosted classifier that has been built up from the previous stages of training. $E(F)$ is some of the error function and $f_t(\mathrm{x}) = \propto_t h(x_i)]$ is the weak learner that is considered for addition to the final classifier.

C. Feature Types

1) Haar Like Feature

Haar-Like Features are rectangular digital images features which are like Haar wavelets, they are using the combination of simple Haar-like features to classify an object. Haar wavelets are functions of calculating the difference in intensities on an object. The function is used in different size regions and different position on the detection window. Those features are defined as the difference of the sum of the pixels within a rectangular region. The feature values are pixel values in the white rectangles subtracted from the black rectangles [27]. Haar-like features can be classified as two, three and four rectangle features as shown in Fig. 2.

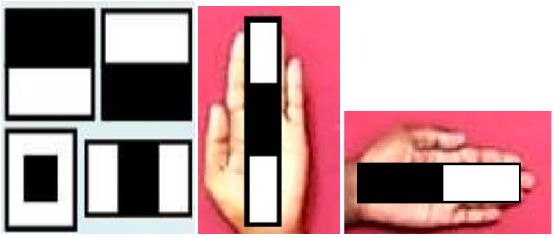


Fig. 2. Haar-Like Features

2) HOG features

HOG refers to Histogram of Orientated Gradients. It is proposed to characterize the local objects appearance and shape using the distribution of local intensity gradients. The overall silhouette of an object is captured by the HOG features. The orientation of HOG is formed by arranging the image gradients magnitude and orientation, like a two-dimensional color histogram. The image is divided into smaller cells as in Fig. 3, and the gradients are calculated from each cell.

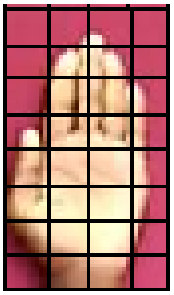


Fig. 3. Division of hand image into cells

3) Local Binary Pattern(LBP) features

LBP was initially proposed for the characterization of the texture of the materials such as wood, cotton, and sand, by modeling their patterns. This is done by analyzing the grayscale values of circularly symmetric neighborhood sets the distance from a given center pixel. The interpolation is then used to calculate the grayscale values of the locations that do not fall at the center of the pixel [12]. LBP operator labels pixels of the input image by thresholding neighboring P pixels with the center pixel value in a circle of radius R, considering the resulting sequence as binary number see Fig. 4.

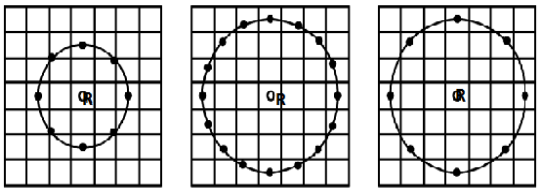


Fig. 4. Circular neighbors set for three different values of P and R

Given a central pixel, the resulting LBP in decimal form is expressed as:

$$LBP_{P,R} = \sum_{k=0}^{P-1} s(\text{ip} - \text{ic})2^p \quad (3)$$

i_c= Grey level intensities of the central

i_p= The surrounding pixel, p running over neighboring pixels

$$s(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases} \quad (4)$$

IV. METHODOLOGY

In Fig.5, the diagram of the methodology is shown. The following steps are presented.

A. Prepare Images

A video of 30 minutes was captured using a cell phone at the resolutions of 1280 x 720. video frames from a captured video were converted into images in jpg format. The images were grouped into positive and negative images for the preparation of training.

B. Crop Image

The region of interest (ROI) with different angles were cropped as shown on Fig.6 from the positive image to create a data file of ROI. The complete image is not provided for feature extraction but some region of interest from the image is selected.

C. The region of Interest (ROI)

300 cropped regions of interest with different angles as shown in Fig. 7, were selected to extract the features of that region for training the classifier. Positive images are images which contain the region of interest, which is the hand.

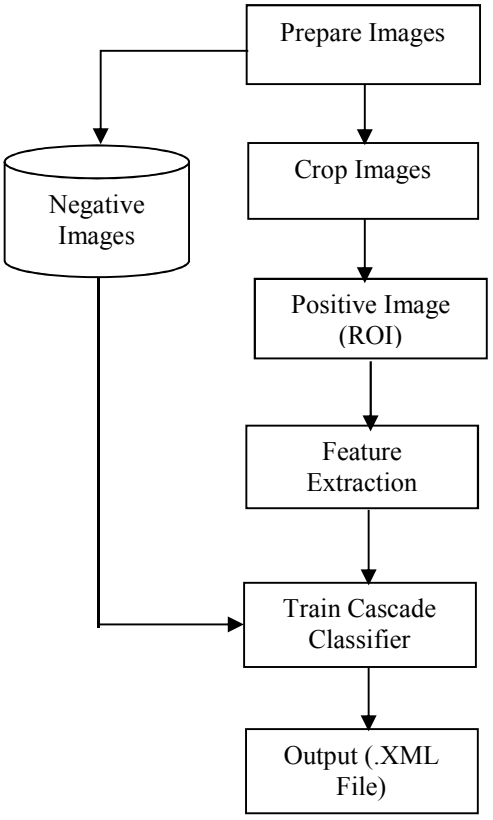


Fig. 5. The block diagram of the methodology

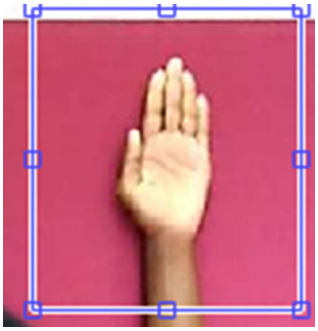


Fig. 6. Cropped image

Fig. 7. Positive images

D. Negative Images

Fig. 8. Negative images

All other pictures which do not contain the region of interest are called negative images, for instance, the images are shown in Fig. 8. In this experiment, 10 000 negative image databases were created for the training purpose.

E. Feature extraction

After the selection of the region of interest (hand), the hand features were extracted using Haar, HOG, and LBP.

F. Train the classifier

The classifier was trained using positive (ROI) and negative images, so that the final classification can be able to differentiate between the hand and the non-hand. Different feature types were trained while the parameters were tuned per feature extraction type.

G. Output

The .XML file was created. The .XML file contained all the information regarding the classifier.

H. Test and record the results

Fig. 9. Detected hand image

The performance of the trained model was tested using the.XML file by providing the test images to the classifier. 320 test images were used to test the three feature detectors. The bounding box is drawn on the detected object or hand as shown in Fig. 9.

V. THE EXPERIMENTAL RESULT AND ANALYSIS

This section shows the experiments validating the effects of features and parameters on the classification performance.

TABLE I. TERMS FOR CLASSIFICATION BY FEATURE DETECTORS

Term	Description	Symbol
True Positive	Successful classification of positive image	TP
False Positive	Misclassification of negative image as positive image	FP
False Negative	Misclassification of positive image as negative image	FN
FalseAlarmRate	Percentage of acceptable false positive per stage	FAR
NumCascadeStages	Number of cascaded stages	NCS
NegativeSampleFactor	The percentage of negative images to be used with respect to positive images used	NSF
TruePositiveRate	Minimum percentage of positive images to be trained per stage	TPR
ObjectTrainingSize	The minimum size of sliding search window (Width "W" and Height "H")	OTS
TotalPositiveSample	Total number of positive images	TPS
NegativeSample	Number of negative images per stage	NS
NumberPositiveSample	Number of positive images per stage	NPS

A. The validating criteria for hand detection performance

The terms are shown in Table I are considered as the main criteria to validate the classifier performance using the three feature detectors and various parameter combinations.

$$NPS = \frac{Total\ Positive\ Sample}{(1+(NCS-1)(1-TPR))} \quad (5)$$

$$NS = NPS * NSF \quad (6)$$

Equation (5) shows how to calculate the number of positive images to be used in each stage and equation (6) indicates how to calculate the number of negative images.

The experiment was done using 300 positive images (ROI), 10 000 negative images and 320 test images which all contain the hand. The sample size is 32x32 pixels.

B. The effects of parameters on the classification performance using selected features

Experiments are designed to validate how the parameters affect the classification performance for the features selected.

1) LBP features

TABLE II. LBP FEATURE PERFORMANCE RESULTS

Training Parameters		**Performance of LBP**					
FAR	***TPR***	***Actual Trained Stages***	***Samples used for training***	***Training time***	***FP***	***FN***	***TP***
0.1	**0.1**	5/5	65	2 sec	33867	2%	98%
0.1	**0.5**	5/5	100	3 sec	28216	31%	69%
0.1	**0.995**	5/5	294	12 sec	11510	5%	95%
0.1	**1**	5/5	300	94 sec	478	15%	85%
0.5	**0.1**	5/5	65	2 sec	33867	1%	99%
0.5	**0.5**	3/5	100	1 sec	27566	6%	94%
0.5	**0.995**	4/5	294	6 sec	23230	26%	74%
0.5	**1**	3/5	300	4 sec	37816	32%	68%
0.1	**0.1**	10/10	32	3 sec	18717	30%	70%
0.1	**0.5**	10/10	54	8 sec	1529	49%	51%
0.1	**0.995**	10/10	287	523 sec	140	26%	74%
0.1	**1**	6/10	300	900 sec	45	31%	69%
0.5	**0.1**	10/10	32	2 sec	18717	30%	70%
0.5	**0.5**	8/10	54	4 sec	12453	38%	63%
0.5	**0.995**	10/10	287	27 sec	12983	6%	94%
0.5	**1**	7/10	300	27 sec	15605	0%	100%

From Table II, the following results are obtained for the parameter combination (training stage=5, FAR=0.1, and TPR varying from 0.1 to 1): (1) all 5 stages are used, (2) the number of samples used for training increases from 65 to 300, (3) the training time increases, (4) the FP is reduced, (5) the FN fluctuates, and the TP also fluctuates. When the FAR increased to 0.5 and the TPR varying from 0.1 to 1, the classifier trained stages fluctuate, the sampled used for training increases from 65 to 300, the time taken to train fluctuates, the FP decreases and increase when TRP is 1, the FN increases and the TP decreases. When the training stages increase to 10, FAR=0.1 and the TPR varying from 0.1 to 1, the classifier trained all stages except when TPR= 1, the samples used for training increases from 32 to 300, the training time increases, the FP decreases, FN fluctuates, and TP as well fluctuates. When the FAR increases to 0.5 and the TPR increases from 0.1 to 1, the classifier trained stages fluctuate, the samples used for training increases from 32 to 300, the training time increases, the FP fluctuates, the FN fluctuates, and the TP fluctuates.

2) HOG features

The results from Table III indicate that when the training stage is set to, the FAR is set to 0.1, and TPR is varying from 0.1 to 1, classifier is able to train all 5 stages, the sample used for training increases from 65 to 300, the training time fluctuates, the FP is increases, the FN fluctuates, and the TP also fluctuates.

TABLE III. HOG PERFORMANCE RESULTS

Training Parameters		**Performance of HOG**					
FAR	***TPR***	***Actual Trained Stages***	***amples used for training***	***Training time***	***FP***	***FN***	***TP***
0.1	**0.1**	5/5	65	65 sec	0	66%	3%
0.1	**0.5**	5/5	100	51 sec	9	92%	8%
0.1	**0.995**	5/5	294	95 sec	240	36%	64%
0.1	**1**	5/5	300	119 sec	472	46%	54%
0.5	**0.1**	2/5	65	1 sec	5139	30%	70%
0.5	**0.5**	2/5	100	1 sec	15516	71%	29%
0.5	**0.995**	5/5	294	2 sec	17492	26%	74%
0.5	**1**	5/5	300	2 sec	15304	34%	66%
0.1	**0.1**	7/10	32	3279 sec	0	94%	6%
0.1	**0.5**	6/10	54	4175 sec	0	100%	0%
0.1	**0.995**	8/10	287	6129 sec	3	66%	34%
0.1	**1**	8/10	300	6013 sec	1	68%	32%
0.5	**0.1**	4/10	32	38 sec	49	72%	28%
0.5	**0.5**	3/10	54	38 sec	5	71%	29%
0.5	**0.995**	9/10	287	61 sec	1745	13%	87%
0.5	**1**	10/10	300	29 sec	1457	4%	96%

When the FAR increased to 0.5 and the TPR varying from 0.1 to 1, the classifier trained stages increases, the sampled used for training increases from 65 to 300, the time taken to train increases, the FP fluctuates, the FN fluctuates, and the TP fluctuates. When the training stages increase to 10, the FAR set to 0.1 and the TPR varying from 0.1 to 1, the classifier trained stages fluctuate, the sample used for training increases from 32 to 300, the training time increases, the FP fluctuates, FN fluctuates, and TP as well fluctuates. When the FAR increases to 0.5 and the TPR increases from 0.1 to 1, the classifier trained stages increases, the samples used for training increases from 32 to 300, the training time fluctuates, the FP fluctuates, the FN decreases and the TP increases.

3) Haar-like features

The results from Table IV indicate that when the training stage is set to, the FAR is set to 0.1, and TPR is varying from 0.1 to 1, classifier trained stages varies, the sample used for training increases from 65 to 300, the training time varies, the FP is varying, the FN fluctuates, and the TP also fluctuates. When the FAR increased to 0.5 and the TPR varying from 0.1 to 1, the classifier trained stages fluctuate, the sampled used for training increases from 65 to 300, the time taken to train fluctuates, the FP fluctuates, the FN fluctuates, and the TP fluctuates. When the training stages increase to 10, the FAR set to 0.1 and the TPR varying from 0.1 to 1, the classifier trained stages fluctuate, the sample used for training increases from 32 to 300, the training time fluctuates, the FP fluctuates, FN fluctuates, and TP as well fluctuates. When the FAR increases to 0.5 and the TPR increases from 0.1 to 1, the classifier trained stages fluctuate, the samples used for training increases from 32 to 300, the training time fluctuates, the FP fluctuates, the FN fluctuates, and the TP fluctuates.

TABLE IV. HAAR_LIKE PERFORMANCE RESULTS

Training Parameters		**Performance of Haar**					
FAR	***TPR***	***Actual Trained Stages***	***amples used for training***	***Training time***	***FP***	***FN***	***TP***
0.1	**0.1**	5/5	65	39 sec	500	51%	49%
0.1	**0.5**	4/5	100	262 sec	363	60%	40%
0.1	**0.995**	5/5	294	677 sec	643	3%	98%
0.1	**1**	5/5	300	494 sec	1301	0%	100%
0.5	0.1	2/5	65	9 sec	17449	1%	99%
0.5	0.5	2/5	100	14 sec	14766	27%	73%
0.5	0.995	5/5	294	242 sec	32060	19%	81%
0.5	1	4/5	300	197 sec	29389	0%	100%
0.1	**0.1**	9/10	32	1527 sec	26	60%	40%
0.1	**0.5**	7/10	54	1119 sec	5	71%	29%
0.1	**0.995**	6/10	287	1556 sec	348	10%	90%
0.1	**1**	7/10	300	1635 sec	302	0%	100%
0.5	0.1	4/10	32	12 sec	1222	7%	93%
0.5	0.5	4/10	54	29 sec	2760	16%	84%
0.5	0.995	9/10	287	515 sec	6934	0%	100%
0.5	1	8/10	300	482 sec	11785	0%	100%

C. The effects of features on the classification performance

The comparisons among the selected features are demonstrated on the bar graphs in Figs. 10-13. With few stages in this instance 5 stages, and low FAR (0.1) the LBP feature detector performed better that Haar and HOG as shown in Fig.10, when the FAR is increased to 0.5, Haar performs better than LBP and HOG, as shown in Fig.11. Experimenting feature detectors with 10 stages, Haar feature detector improves its performance especially when the TPR increases towards 1, see Fig. 12 and 13, and all the feature detectors performs better with FAR equals to 0.5, also when

the training stages increases as well as the TPR increases towards 1 as shown in Fig. 13. The experiment indicates that the more you increase the number of training stages, the FAR to 0.5 and the TPR towards 1 the feature detectors performs much better than when they are low. It further indicates that the training time is longer especially with Haar but it out performs the remaining two feature detectors.

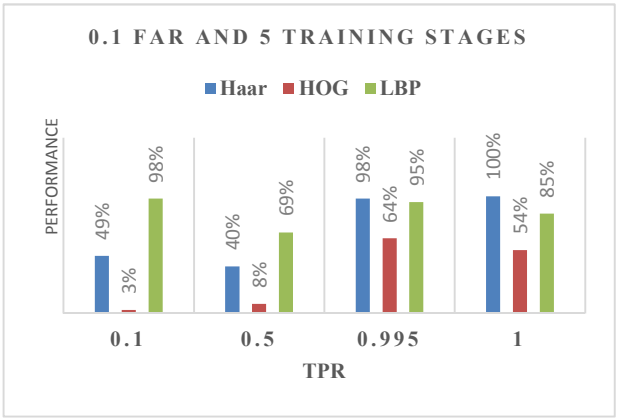


Fig. 10. Classification rate comparison (FAR=0.1, Stages=5)

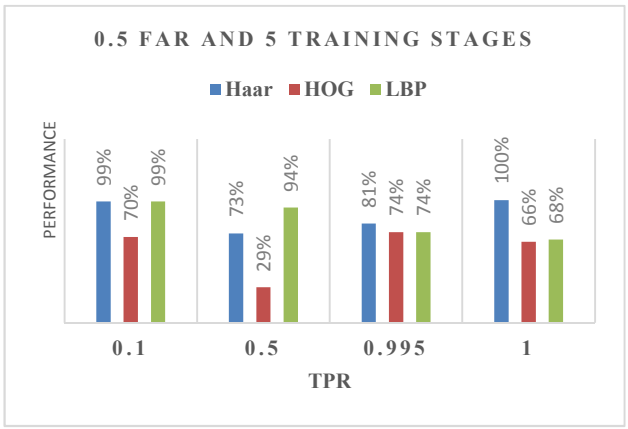


Fig. 11. Classification rate comparison (FAR=0.5, Stages=5)

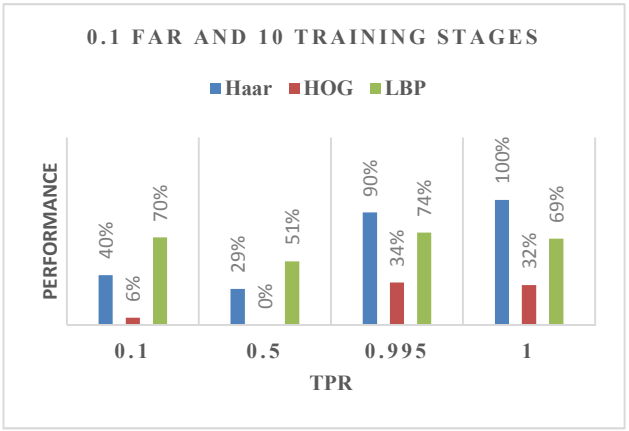


Fig. 12. Classification rate comparison (FAR=0.1, Stages=10)

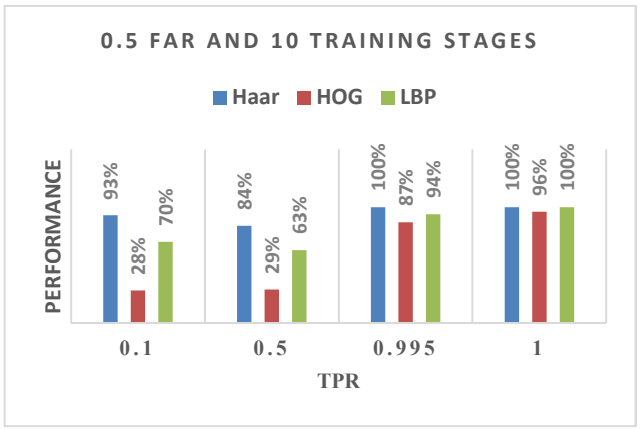


Fig. 13. Classification rate comparison (FAR=0.5, Stages=10)

Various parameter combinations were considered to validate the effectiveness of the selected features as shown in figure 10 to 13. The AdaBoost cascade classifier was employed as the test bed. From the experiment results and analysis, it is noted that: (1) FalseAlarmRate, TruePositiveRate and training stages have significant effects on the classification performance for all features. For instance, to obtain the higher detection using any of the feature (Haar, LBP, or HOG), FalseAlarmRate has the favorable value of 50%, and the TruePositiveRate should have the value close to 100% and train as many stages as possible. (2) The FP is opposite to FN i.e when the FN is at the minimum the FP is at maximum. (3) When the TP is at maximum the FN is at the minimum. When the TP is at maximum the FP is at maximum. With the same parameter setting: (1) HOG intending to get lower FP on the cost of lower TP. (2) Provided reasonable ROI can be determined before classification (such as using tracking techniques), LBP features is more competitive with lower FAR. (3) Haar-like features win for higher FAR. Figure 10 to 13 indicate the whole explanation with different parameters tuned.

VI. CONCLUSION

This experiment was conducted to test three different feature detectors i.e. Haar-Like features, HOG and LBP on detecting the palm of the hand on a constant background. A high-resolution camera was used to capture three different types of images i.e. positive, negative and test images to fulfil the comparison experiment. The region of interest was cropped to make the positive images for training. The Viola and Jones method of cascade adaptive boost classifier was used to train the three feature detectors. 10 620 images were used for the experiment, where 10 000 were negative images, 300 were positive images and 320 were for the testing of the classifier. The reason for the number of images used was only to be able to train more stages, there is no other special reason. The three feature detectors were trained one after another using the same computer and number of images. Different parameters were turned during training. The different feature detectors were tested individually using 320 test images. All the three feature detectors were compared based on various parameter settings of TPR, and FAR. The performance was looked based on number of training stages, samples used for training, time taken for training, false

positive, false negative and true positive. The experiment indicated that the more you increase the number of training stages, and also increasing the TPR as well as FAR the feature detectors performs better, however in this experiment Haar-Like features performs better than HOG and LBP.

ACKNOWLEDGMENT

This work was supported in part by the National Research Foundation of South Africa (Grant no. 92468 and 93539).

REFERENCES

[1] G.S. Sujatha, and V.V. Kumari, An innovative moving object detection and tracking system by using a modified region growing algorithm, vol 7(2), pp 39-55. Signal and Image Processing International Journal, April 2016.

[2] E. Baser, and Y. Altun, Detection, and classification of vehicles in traffic by using haar cascade classifier, vol 4(2), pp 137-140. International Journal of Advanced Electronic and Computer Science February 2017.

[3] Y. Tang, Y. Tang, Y.C. Xu, and C.Z. Zhang, Bobust vehicle detection based on cascade classifier in traffic surveillance system, vol 6, pp 349-354. The Open Automation and Control System Journal, 2014.

[4] Z. Xia, C. Yuang, X. Sun, D. Sun, and R. Lv, Combining wavelet transform and LBP related feature for fingerprint liveness detection, vol 43(3). IAENG International Journal of Computer Science, 27 August 2017.

[5] M. Oualla, A. Sadio, and S. Mbarki, A new approach to represent rotated Haar-Like features for objects detection, vol 78(1), pp 15-23. Journal of Theoretical and Applied Information Technology, 10 August 2015.

[6] A. Bansal, R. Agarwal, and R.K. Sharma, Statistical feature extraction based iris recognition system, vol 41(5), pp 507-518. Indian Academy of Science, May 2016.

[7] P.A. Viola, and M. Jones, Rapid object detection using a boosted cascade of simple features, Conference on Computer Vision and Pattern Recognition, pp 1 - 9, 2001.

[8] O.M. Patil, and V. Gaikwad, Hand gesture recognition using Haar-Like Features and cascade classifier, vol 6(2), pp 1-6. International Journal of Advanced Research in Computer Science and Management Studies, February 2018.

[9] N. Dala, and B. Triggs, Histogram of Orientated Gradients for human detection, Conference of Computer vision and pattern recognition, vol 1, pp 886-893, 2005.

[10] M.E. Yildirim, J.S. Park, J. Song, and B.W. Yoon, Gender classification based on binary Haar cascade, vol 3(2), pp 105-107. International Journal of Computer and Communication Engineering, March 2014.

[11] M. Wisim, A. Aziz, and S.F. Ali, Object's shape recognition using Local Binary Patterns, vol 8(8), pp 258-262. International Journal of Advanced Computer Science and Application, 2017.

[12] A. Rahim, N.l Hossain, T. Wahid, and S. Azam, Face recognition using Local Binary Patterns (LBP), vol 13(4). Global Journal of Computer Science and Technology Graphics and Vision, 2013.

[13] E. Alionte, and C. Lazar, A practical implementation of face detection by using Matlab cascade object detector, 19th International Conference on System Theory, Control and Computing, pp 785-790. Cheil Gradistei, Romania, 14-16 October 2015.

[14] T.T. Do, and E. Kijak, Face recognition using co-occurrence Histogram of Oriented Gradient, 37th International Conference on Acoustics, Speech, and Signal Processing, pp 1301-1304. March 2012.

[15] C. Ballur, and S.S. Shylaja, Application of local binary pattern and principal component analysis for face recognition, vol 1, Pp 34-39. International Journal of Electrical, Electronics and Data Communication, July 2013.

[16] S. Paul, S. Dixit, and C. Dabas, Object detection for markerless augmentation using Haar training, vol 6(1), pp 43-46. International Journal of Advanced Research in Computer Science, January to February 2015.

[17] T.A. Al-asadi, and A.J. Obaid 2, Object detection and recognition by using the enhanced speed up robust feature, vol 16(4), pp 66 – 71. International Journal of Computer Science and Network Security, April 2016.

[18] N. Adinarayana, B.A. Reddy, and G.B. Reddy, Comparative studies of feature extraction in face recognition system by using Haar wavelet transform, vol 3, pp 60-63. International Conference on Advances in Computing, Electrical and Communication Engineering, May 2017.

[19] J. Vivek, A. Binu, and A.J. Kuttyamma, Licence plate detection and recognition using a HOG feature based SVM classifier, vol 5(5), pp 10115-10119. International Journal of Innovative Research in Computer and Communication Engineering, May 2017.

[20] D Anh, S. Pao, and K. Wataru, Fish detection by LBP cascade classifier with optimized processing pipeline, vol 82(9), pp 1-4. Information Processing Society of Japan SIG Technical Report, 2013/9/12.

[21] S. Asnani, S.D. Waseem, and A.A. Manjotho, Unconcealed gun detection using Haar-like and HOG features – A comparative approach, vol 4(1), pp 34-39. Asian Journal of Engineering, Science and Technology, March 2014.

[22] R.L. Kulkarni, Handwritten character recognition using HOG, COM by OpenCV and Python, vol 5(4), pp 36-40. International Journal of Research in Computer Science and Management Studies, April 2017.

[23] J.E.C. Cruz, E.H. Shiguemori, and L.N.F. Guimaraes, A comparison of Haar-like, LBP, and HOG approaches to concrete and asphalt runaway detection in high-resolution imagery, vol 5(4), pp 121-136. Journal of Computational Interdisciplinary Science, 2015.

[24] D.K. Shedge, B. Mathai, P. Barude, S. Katkamwar, and A. Ranjan, Road sign detection using cascade classifier, vol 8(4), pp 979-983. International Journal of Science and Engineering Research, April 2017.

[25] V. Singh, V. Shokeen, and B. Singh, Face detection by Haar cascade classifier with simple and complex backgrounds images using OpenCV implementation, vol 1(12), pp 33 – 38. International Journal of Advanced Technology in Engineering and Science, December 2013.

[26] R. Kaur, and V. Chopra, Implementing AdaBoost and enhanced AdaBoost algorithm in web mining, vol 4(7), pp 306-311. International Journal of Advanced Research in Computer and Communication Engineering, July 2015.

[27] C.C. Cia, J. Gao, B. Minjie, P. Zhang, and H. Gao, Fast pedestrian detection with an AdaBoost algorithm using GPU, vol 8(6), pp 125-132. International Journal of Database Theory and Application, 2015.

Evaluating the Effects of Hardware Configurations on Bro under DDoS Attacks

Welcome Luthuli, Olukayode Oki, Paul Tarwireyi and Matthew Adigun
Department of Computer Science
University of Zululand
KwaDlangezwa, RSA
okikayode@gmal.com

***Abstract*—The exponential growth of network traffic and the growing sophistication of network attacks call for faster, efficient and scalable intrusion detection systems (IDS) that will be able to quickly look into the traffic and timeously produce alerts when malicious traffic has been detected. Snort has been the de-facto standard of IDS for so many years but has been ineffective under heavy loads. In both multi-core and single core hardware configuration snort shows no improvement in detection capability of Transmission Control Protocol (TCP) flooding Distributed Denial of Service (DDoS) attack. This has led to the development of alternative IDS which try to address the limitations of Snort. Bro is a flexible script-driven intrusion detection system, which provides a 'worker' based architecture to utilize multiple processors. The aim of this paper is to evaluate Bro in terms of performance and packet handling against TCP flooding DDoS attacks under different hardware configurations. To achieve this aim, Bro was installed on different hardware configurations. Tests were conducted to assess its performance under each configuration. Packet loss, Throughput and resource utilization metrics were measured. The results show that utilizing better hardware increases resources availability hence gives the system better performance.**

Keywords—Bro, IDS, DDoS, Hping3, Ostinato

I. INTRODUCTION

Security in wireless network communication is an important problem, which has been there for a long time. Security usually follows a multi-layer approach, also known as defense-in-depth. Defense-in-depth is the concept of protecting a computer network with a series of defensive mechanisms such that if one mechanism fails, another would already be in place to stop the attack [1]. The Intrusion Prevention Systems (IPS), which consists of preventive measures are usually considered as the first layer in network security. Intrusion prevention is a pre-emptive approach to network security used to identify potential threats and respond to them swiftly. The second layer of the network security measure is the Intrusion Detection Systems (IDS) which consists of detective measures. IDS monitors network and analyze information to detect possible security breaches and anomalies.

Although, the network administrators monitor the network but there is a need for more efficient methods to detect and possibly prevent the attacks. Hence, intrusion detection/prevention systems are usually deployed in a network to detect and prevent network security threats. Snort has been the de-facto standard for many years. There are two types of IDS, Knowledge-based (Signature-based) and Behaviour-based (Anomaly-based) IDS. In Knowledge-based IDS, threats are identified by matching each input event against pre-defined signatures that model malicious activity, which means that these types of IDS only detect threats that are already known. While the Anomaly-based IDS monitors the network communication by searching for anomalous behavior and generate alerts.

Studies have been conducted on snort to determine its performance under different hardware configuration [2]. The achieved results from those existing studies [2-5] shows that, the packet handling capacity of snort can be increased by utilizing better hardware. Snort uses packet-based architecture, hence it is ineffective under heavy loads. Also, because of the signature-base of snort, it lacks the ability to detect attacks that do not match any signature. An example of attack that snort might not detect is the flooding DDoS attacks, which do not match with any signature in particular. Hence, there is need for a system that can detect any anomalous behaviour in the network.

This paper looks at Bro as a network security monitor. Bro is a passive IDS and is anomaly based. In deploying an intrusion detection system, a lot is to be taking into consideration, as part of that we need to know the maximum capacity that a chosen IDS can operate. In order to form a proper security layering, we need to know the requirements at which each system works bests. There are a number of intrusion detection systems such as snort and Suricata. Bro is event based, that in combination with it being anomaly-based gives Bro advantage over its peers. Bro categorizes all network activity, reports what is bad and gives the overall activities of the network. Bro users are not only limited to the default functions they can write programs on top of Bro using the scripting language. Unlike Snort or Suricata, Bro does not offer inline intrusion prevention features; it offers new features through its script decision options to drop, sample, throttle, or redirect packets [3].

In this paper, we evaluate the effects of Bro against the distributed denial of service. Hardware configurations were varied to determine their impacts on the performance of the

evaluated Bro. the evaluation was conducted using three metrics; CPU usages, memory and packet loss.

The remainder of this paper is organized as follows: Section II presents a review of existing research work in the area of network security. Section III discusses the evaluation methodology employed in this paper, while Section IV presents the performance evaluation results. The paper is concluded in Section V with an outline of the future work

II. LITERATURE REVIEW

In [4], Srivastava et al conducted a comparative study of various traffic generator tools. The traffic generating tools considered in their study are PackETH, Ostinato, and D-ITH. The comparison was conducted using both TCP and UDP traffic. Their results show that for UDP traffic, the throughput of 9Gbps is achieved by both PackETH and Ostinato, using 5120 and 8195 Bytes respectively.

Three intrusion detection systems under various attacks and rule sets were evaluated in [5]. The evaluated IDS include: Bro, snort, and suricata. The experiments were conducted using different traffic rates and set of active rules. Eight types of attacks were used for this study. The metrics measured in their study include; CPU usage, number of packet loss, and the number of alerts. The result of their experiments showed that increasing the traffic rate affects the performance of IDS in terms of CPU usage, number of packet loss, and alerts. However, their study did not consider the hardware configurations, which has effect on some of their achieved results.

Sommers, et al. in [6] describe a traffic generation framework for conducting online evaluations of network intrusion detection systems over a wide range of realistic conditions. The NIDS evaluated in their study were Bro and Snort. The first set of experiments were designed to establish a baseline of alarm behaviour for each NIDS. In the second set of experiments, the mix of flow volumes between the data traffic generated was altered by Harpoon, and malicious traffic generated by MACE (CSL) or attack-replay (DARPA). In the third set of experiments, they used three different levels of traffic volumes. The results of their study shows that the content, mix, and volume have a tremendous effect on NIDS performance. The DARPA IDS evaluation dataset has been criticized and considered by many as a very outdated dataset, unable to accommodate the latest trend in attacks [3]. In study [6], the version of Bro used is outdated, and hardware specifications were not considered as a factor that can affect the performance of intrusion detection systems.

In [2], Snort was evaluated using different hardware configurations. Like all signature-based NIDS, snort also lacks the ability to detect attacks that do not match with any of the signatures. Flooding DDoS attack does not match with any signature. Three test benches were used for their evaluation, in combination with different hardware configurations. In their experimental setup, the network consists of four computers, in which there is a victim computer, snort system (IDS mode), background traffic generating system, and an attacker system. All these computers are connected using a Cisco 2960 series switch. To simulate an attack, hping3 tool was used, and for background traffic, a tool called ostinato was used together with hping3. Evaluation metrics for their study are maximum packet rate, resource availability (CPU and memory), throughput, and evaluation of snort rate filter option. Based on their achieved results, it was shown that the packet handling capacity of snort can be increased by utilizing the better hardware configurations. Also, it was shown that the multi-core processors do not improve packet-handling capacity and that Snort is ineffective under heavy loads. Hence, there is need for the evaluation of an IDS that handles heavy traffic under different hardware configurations.

III. EVALUATION METHODOLOGY

The purpose of this study is to evaluate the performance of Bro in terms of its ability to handle TCP flooding DDoS attacks when subjected to different hardware configurations. The evaluation is conducted using test benches, while the simulation is done for attack and background traffic.

a) Test benches

To test the performance of Bro, test benches comprising different hardware configuration has been selected. The summary of different hardware configurations used in our test benches are presented in table 1. These test benches are physical machines since NIDS show some limitations on virtual platforms [7]. Each bench consists of four computers with Linux operating system.

Using the setup depicted in Figure 1 with the installed tools presented in table 2, two types of data traffic were generated; background traffic (UDP and ICMP) and the attack traffic (TCP flood DDoS).

TABLE 1: HARDWARE CONFIGURATION FOR EACH TEST

Test bench no.	**System Specifications**			
	Processor	**CPU (s)**	**RAM**	**OS**
1.	Pentium(R) Dual-core @ 2.50 GHz	2	4 GB	Ubuntu 16.04
2.	Pentium(R) Dual-core @ 3.20 GHz		2 GB	
3.	Intel® Core™ i5-6500 CPU @ 3.20GHz	4	8 GB	
4.	Intel® Core™ i7-6700 CPU @ 3.40GHz	8	8GB	

TABLE 2: TOOLS SPECIFICATION FOR EACH SYSTEM

System	**Installed Tools on System**
Victim System (PC 1)	Gnome System Monitor
Bro version 2.5.1 (PC 2)	Gnome System Monitor Wireshark
Background Traffic Generating System (PC 3)	Hping3, Ostinato
Attacker System (PC4)	Hping3
Switch	Catalyst 2950 Switch (24) 10/100 Ethernet ports

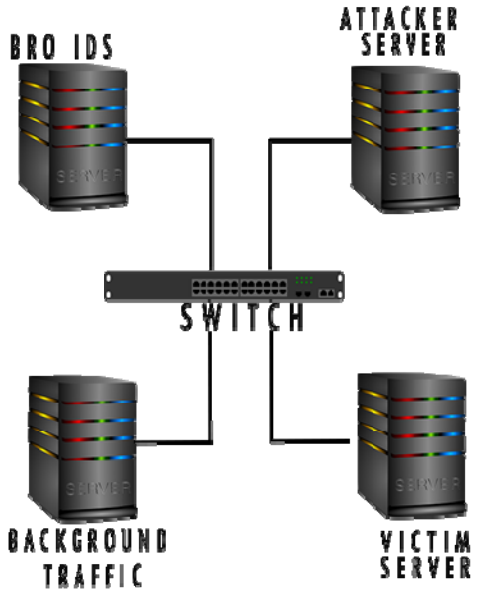


Figure 1. Network diagram

The aim was to investigate the performance of the server per unit time. To simulate the attack, Hping3 has been chosen for its ability to send custom TCP/IP packets and to display target replies similar to that of ping with ICMP replies. Hping3 can handle fragmentation, and almost arbitrary packet size and content, using the command line interface [8], which makes it easy to simulate an attack. Background traffic was generated using Ostinato, which is a packet crafter, network traffic generator and analyser that crafts and send packets of several streams with different protocols at different rates [9]. The rate of packet loss of background data traffic were calculated using Wireshark. In order to determine the performance of Bro, a monitoring Top command from Ubuntu package were used. Top is a terminal command that collects and reports system activity information such as CPU and memory utilization.

IV. Evaluation Metrics

- *Packet Loss:* The packet loss metric is defined as the average number of packet that were lost from the background traffic before reaching their intended destination.

- *Resource Availability:* The aim of DDoS is usually to deplete resources of a system, such that the resources will no longer be available to legitimate users. The resource used by Bro can be used as an indicator of system availability. For the system resources metric, we considered both the CPU and memory utilization. To keep track of the resources, a Top command was used, which automatically capture data at every one minute intervals. The interval will allow the stability of the system before taking results.

- *Average Throughput:* This metric measure the number of packets sent by legitimate clients and successfully received by the server per unit time. It reports the number of packet the system can handle.

Traffic Scenarios:

- Scenario 1(Attack only): The attack is simulated using Hpng3
  Command: sudo hping3 –d 120 –S –w 64 –flood –rand-source <ip addr>

One of the objectives of evaluating Bro, was to investigate its resource utilization. The attack is sent to the victim with no other traffic on the network, to determine the resource utilization of the attack traffic alone. The attack traffic simulates an intrusion, under the attack traffic the CPU and memory utilization were measured to determine the resource utilization under attack.

- Scenario 2 (Mixed Traffic): in a real-world network deployment, at any given time there is traffic going through the network, hence an attack can occur while the network is operational. Mixed Traffic is the combination of attack and background traffic. The resource utilization is closely coupled with packet handling. The number of packets dropped measures packet handling. While DDoS aim at exhausting the host resources such that, the legitimate packets users such as background traffic cannot gain access into the resources.

- Scenario 3 (Background Traffic): The background traffic is the benign traffic that is flowing in the network. For the evaluation of IDS, the knowledge of the system under normal traffic needs to be known, so as to determine the effect of the attack. The background traffic was generated using Ostinato, and it consists of UDP and ICMP traffic.

V. Evaluation Results

The results of the performance evaluation of Bro against DDoS attacks when subjected to different hardware configurations are presented in this section. Each of the reported graphs is the result of the average of ten experiments for each setup considered.

1. *Resource availability*

Bro is a single threaded which means it only uses one core of the CPU. Figure 2 depict the CPU usage under background traffic. The background traffic is constant in terms of volume

and rate, hence Bro does not use much resources in monitoring this traffic because the traffic behavior is normal.

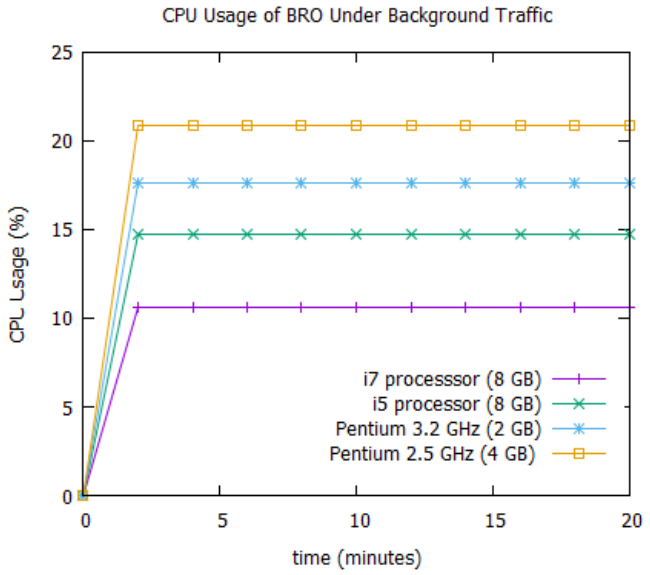


Figure 2. CPU usage under normal network activities

Figures 3 and 4 show that the CPU reaches 100 % usage within a short period of time, this is because TCP flood traffic behaves abnormally, since it changes behavior frequently. The CPU usage of Pentium 3.2 GHz reaches 100 % in a relatively longer time, this is because Pentium 3.2 GHz computer has 2 GB of RAM with a 3.2 GHz Processor. The processor has a faster clock speed and low RAM size, hence the CPU does not use its maximum power to process the instructions available on the RAM since the data on the RAM is not enough.

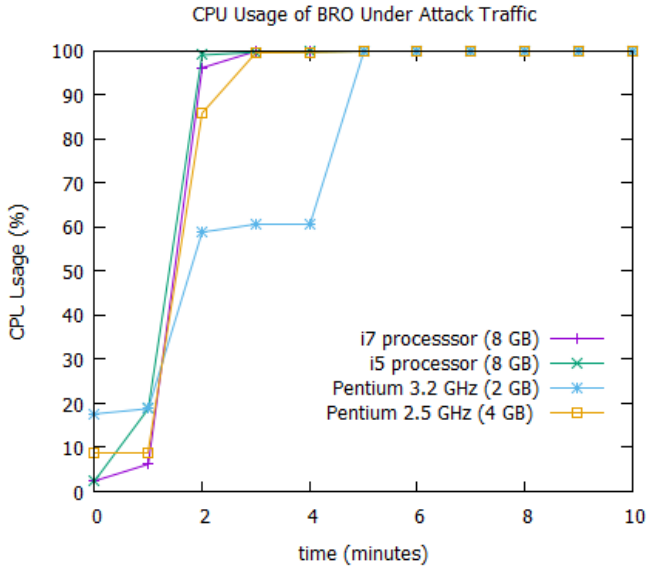


Figure 3. CPU usage under TCP flood DDoS

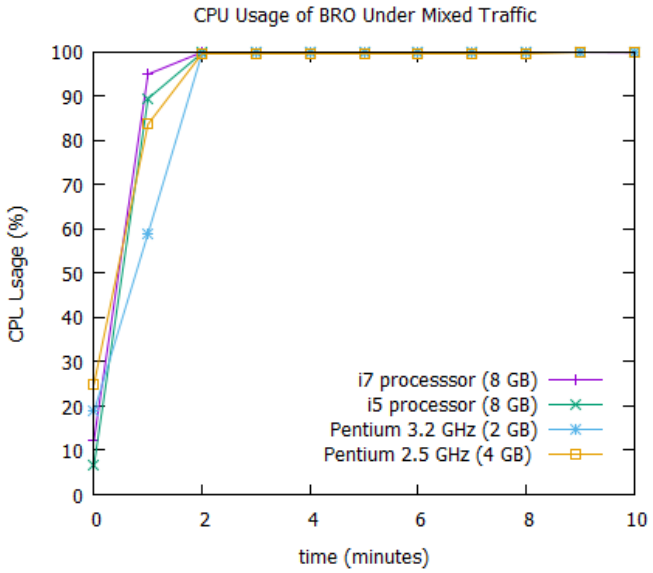


Figure 4. CPU Usage under mixed traffic

The background traffic are UDP and ICMP. UDP traffic were sent with a fixed payload, and that payload consumes a significant space in the system memory especially when using a machine with a low memory. The attacks were sent using flood, which sends as many packets as possible depending on the resources. Figure 5 depict the memory usage of Bro under benign traffic. The background traffic were sent at 630pps for UDP and 325pps for ICMP which is slower than a flood. The traffic was at a constant rate and volume. It can be observed that Bro does not use lots of memory in monitoring constant traffic. Hence, this makes the memory usage to be at a constant rate.

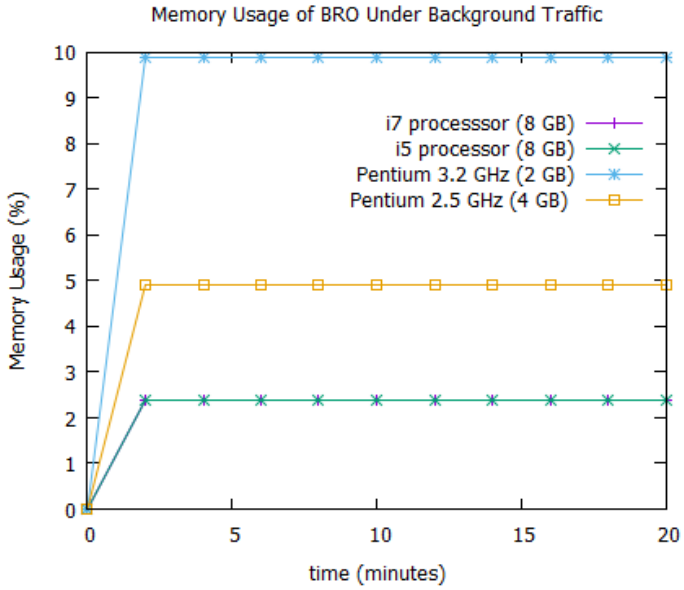


Figure 5. Memory usage under normal traffic

Figures 6 and 7 show that the computer with i7 processor have the highest memory usage under attack and mixed traffic than other computer processors considered. This performance can be attributed to the fact that the i7 computer has more resources in terms of memory and processing power compare to others. The high processing power of i7 computer, helps the system to process more instructions per unit time compared to other machines. Bro takes advantage of system configurations, in which, as the processing power increases more packets are being processed. Hence, more memory were used by Bro in order to process the incoming packets.

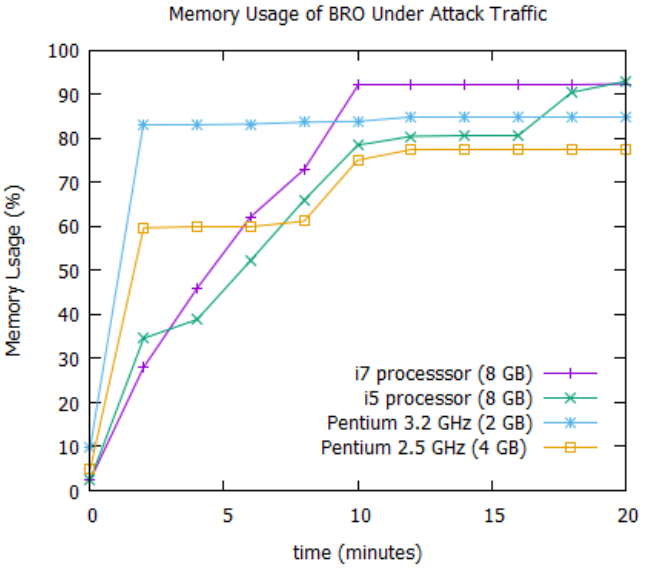


Figure 6. Memory usage under TCP flood traffic

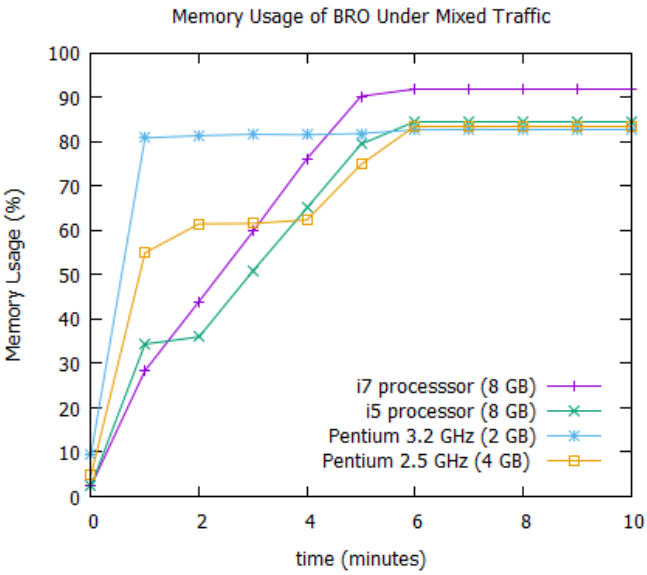


Figure 7. Memory usage under mixed traffic

2. Packet Drop

The packet drop indicates the capability of Bro in handling packets under attack. The background traffic packets were not considered since it is the attack. Hence, only the mixed traffic-based packets were considered. The lower the packet drop the higher the packet handling capability and the better the system performance.

The packet drop is one of the measurements of Bro's performance, the CPU and Memory used by Bro leads to packet drop. The number of packets sent, and the number of packets received for both UDP and ICMP mixed traffic are presented in tables 3 and 4 respectively. It can be observed for both the UDP and ICMP mixed traffic that, as the hardware configurations increases, the packet drop decreases.

TABLE 3: PACKET DROP OF UDP UNDER MIXED TRAFFIC

Machine	Packet sent	Packet received
Pentium(R) Dual-core E5200	378000	28984
Pentium(R) Dual-core E6700	378000	20801
Intel® Core™ i5-6500	378000	90720
Intel® Core™ i7-6700	378000	112824

TABLE 4: PACKET DROP OF ICMP UNDER MIXED TRAFFIC

Machine	Packet sent	Packet received
Pentium(R) Dual-core E5200	195000	19304
Pentium(R) Dual-core E6700	195000	16338
Intel® Core™ i5-6500	195000	72150
Intel® Core™ i7-6700	195000	117000

3. Average Network Throughput

The average Throughput is the number of legitimate packets sent and received by the server over a period of time. The higher throughput values indicate that the system can process more packets per given time. It was observed that there is a tradeoff between the achieved throughput and the packet loss. As the packet loss decreases, the average throughput increases. Tables 5 and 6 present the average throughput for both the UDP and ICMP mixed traffic respectively. It was observed that the computer with i7 processor has the highest throughput. This achieved result by i7 computer can be attributed to the memory and CPU of the machine, both of which control the computer operations. Bro takes advantage of these resources. Hence, the higher the system configurations (processor speed) the better the performance of Bro.

TABLE 5: UDP THROUGHPUT UNDER MIXED TRAFFIC

Machine	Packet sent	Packet received	Throughput (pps)
Pentium(R) Dual-core E5200	195000	19304	32
Pentium(R) Dual-core E6700	195000	16338	27
Intel® Core™ i5-6500	195000	72150	120
Intel® Core™ i7-6700	195000	117000	195

TABLE 6: ICMP THROUGHPUT UNDER MIXED TRAFFIC

Machine	Packet sent	Packet received	Throughput (pps)
Pentium(R) Dual-core E5200	378000	28984	48
Pentium(R) Dual-core E6700	378000	20801	35
Intel® Core™ i5-6500	378000	90720	151
Intel® Core™ i7-6700	378000	112824	188

VI. CONCLUSIONS

The purpose of this study was to evaluate the effect of hardware configurations on Bro under the DDoS attack. The evaluation was conducted using three performance metrics; Resource availability, Packet drop, and average network throughput. The higher the CPU processing power the faster it fetches data from the RAM, which means it will make room for more packets to be processed. Bro categorizes TCP flood as a threat hence it keeps it in a file log. The higher the TCP flood traffic processed, the higher the memory usage. TCP flood uses random source IP address and random flags, which Bro categorizes as a threat pattern. TCP flood is sent at a very high rate which forces Bro to process packets at a high rate. However, been a single threaded process, makes Bro to be over worked under attack.

Packet Drop is due to resource utilization, the more resources are being used up the higher the packet drop. To manage a large volume of traffic, high memory and processing power is needed. The higher the hardware specification the higher the Throughput. Using better hardware improves the performance of Bro, better packet delivery and lower Packet Drop. The resource utilization still needs to be improved. Our future focus is to develop a new version of Bro such that it multithreaded so as to be able to take the advantages of modern computers, which are multithreaded.

REFERENCES

[1] T. McGuiness, "Defense in depth," SANS Institute InfoSec Reading Room, 2001.

[2] A. Saboor, M. Akhlaq and B. Aslam, "Experimental evaluation of Snort against DDoS attacks under different hardware configurations," *Information Assurance (NCIA), 2013 2nd National Conference,* pp. 31-37, December 2013.

[3] G. Khalil, "Open Source IDS High Performance Shootout," SANS Institute Reading Room , 2015.

[4] S. Srivastava, S. Anmulwar, A. M. Sapkal, T. Batra, A. K. Gupta and V. Kumar, "Comparative study of various traffic generator tools," *Engineering and Computational Sciences (RAECS), 2014 Recent Advances,* pp. 1-6, 2014.

[5] K. Thongkanchorn, S. Ngamsuriyaroj and V. Visoottiviseth, "Evaluation studies of three intrusion detection systems under various attacks and rule sets.," *TENCON 2013-2013 IEEE Region 10 Conference (31194),* pp. 1-4, 2013.

[6] J. Sommers, V. Yegneswaran, P. Barford, "Toward comprehensive traffic generation for online IDS evaluation," Technical Report, University of Wisconsin Madison 2005.

[7] M. Akhlaq, F. Alserhani, I.U. Awan, J. Mellor, A.J. Cullen, P. Mirchandani, "Virtualization Efficacy for Network Intrusion Detection Systems in High Speed Environment," *International Conference on Information Security and Digital Forensics,* vol. 41, pp. 26-41, 2009.

[8] S. Sanfilippo, "Hping3 (8)-Linux Man Page," 2005. [Online]. Available: https://linux.die.net/man/8/hping3. [Accessed 10 November 2017].

[9] P. Srivats, "Ostinato packet/traffic generator and analyzer.," 2011. [Online]. Available: www.ostinato.org. [Accessed 15 November 2017].

[10] P. Mehra, "A brief study and comparison of snort and bro open source network intrusion detection systems," *International Journal of Advanced Research in Computer and Communication Engineering,* vol. 1, no. 6, pp. 383-386, 2012.

[11] K.K. Oo, K.Z. Ye, H. Tun, K.Z. Lin, E.M. Portnov, "Enhancement of Preventing Application Layer Based on DDOS Attacks by Using Hidden Semi-Markov Model," *In Genetic and Evolutionary Computing,* vol. 387, pp. 125-135, 2016

[12] C. Thomas, N. Balakrishnan, "Usefulness of DARPA dataset for intrusion detection system evaluation," *Proc. SPIE Int. Defense and Security Symp.* 2008.

[13] A. K. Kyaw, Yuzhu Chen and J. Joseph, "Pi-IDS: evaluation of open-source intrusion detection systems on Raspberry Pi 2," *Information Security and Cyber Forensics (InfoSec), 2015 Second International Conference,* pp. 165-170, 2015

[14] W. J. Buchanan, F. Flandrin, R. Macfarlane, and J. Graves, "A methodology to evaluate rate-based intrusion prevention system against distributed denial-of-service (DDoS)," 2011.

[15] N. J. Puketza, K. Zhang, M. Chung, B. Mukherjee and R. A. Olsson, "A methodology for testing intrusion detection systems," *IEEE Transactions on Software Engineering,* vol. 22, no. 10, pp. 719-729, 1996.

Setting Heuristics for Eye-Tracking Assessment on Divisibility Rules

Pieter Potgieter
Department of Information Technology
Central University of Technology Free State
Bloemfontein, South Africa
pieter@cut.ac.za

Pieter Blignaut
Department of Computer Science and Informatics
University of the Free State
Bloemfontein, South Africa
BlignautPJ@ufs.ac.za

***Abstract*—True/false questions are one of the basic question types widely used in assessments. Although true/false questions are easy to construct and take little time to be answered, there is a high probability of learners guessing the correct answer. A correct answer to a true/false question is not an indication that a learner fully understands the concept being tested. That said, an incorrect answer does not necessarily mean that the learner does not understand the concept being tested. There may be another reason why the learner gave an incorrect answer. Specifically, when a teacher uses true/false questions to assess learners on divisibility rules, the teacher has to ensure that learners use their knowledge to determine the answer. Therefore, careful compilation of the dividends for an assessment can contribute to determining whether the learner applied the divisibility rules correctly. Eye-tracking can be used to provide objective and quantitative information about the user's intended gaze-patterns over a specific stimulus. Therefore, the gaze behaviour of learners can be used as an indication whether they inspected the appropriate digits of the dividend. It is suggested that (i) heuristics on compiling appropriate dividends, and (ii) setting minimum gaze requirements for attention levels per digit of the dividend for each divisor, could assist teachers in compiling effective true/false assessments to test learners' knowledge of divisibility rules.**

Keywords—heuristics, eye-tracking, dividends, divisibility rules, Mathematics

I. INTRODUCTION

A basic and widely used question format is that of multiple choice and true/false questions [1]. There is a high probability that learners could guess true/false questions correctly [2, 3] and learners can pass an assessment regardless of their ability to solve the problem. A teacher can improve the credibility of true/false questions by asking learners to provide reasons for their answers [1]. Unfortunately, there are usually not enough time during an assessment to expect learners to write down motivations for their responses.

Eye-tracking can be used to identify the strategies that participants use to solve mathematical problems, especially where other methods, such as verbal responses, are less reliable [4].

This paper focuses on firstly setting heuristics to compile dividends for assessing the divisibility rules such that learners, who do not know the divisibility rules, would probably give incorrect answers. Secondly, this study aims to define minimum gaze requirements for attention levels per digit of the dividend for each divisor to evaluate whether learners applied the divisibility rules correctly.

The background of divisibility rules, distractors and incorrect reasoning by learners are provided in the next section. This will be followed by the experimental details, results and finally, the proposed heuristics.

II. BACKGROUND

A. Divisibility rules

If learners inspect only the ones digit (the last digit of the dividend) to identify if the dividend is divisible by 2, 5 or 10, they apply the tests for divisibility for these divisors. A dividend is divisible by 4 if the number formed by the last two digits is divisible by 4, and a dividend is divisible by 8 if the number formed by the last three digits is divisible by 8. A dividend is divisible by 3 if the sum of its digits is divisible by 3, and a dividend is divisible by 9 if the sum of its digits is divisible by 9 [5]. A dividend is divisible by 6 if the dividend is divisible by 2 and 3. The number of digits in the dividend has no effect on the divisibility rule.

B. Distractors

The probability that learners can correctly guess answers should be limited as far as possible. Distractors, also called foils or misleads, are incorrect options used to identify learners who do not know how to arrive at the correct answer [6]. Learners who are not aware of the correct divisibility rules may be "guided" by a distractor and therefore base their answer on the given distractor [6]. Therefore, the distractors should be compiled in such a way that learners would find it difficult to guess the divisibility, but easy to perform division without using a calculator [7]. For example, for the divisibility of 75133 by 3, it is difficult to guess the correct answer, but easy to calculate. Since 7+5+1+3+3=19, and 3 does not divide into 19, 75133 is not divisible by 3.

C. Incorrect reasons for answers

Learners sometimes pretend to know the correct strategy to solve a problem, although there is no evidence that they do [8]. Learners may be confident about their reasoning even though it is incorrect [9, 10]. These incorrect reasons will be used to set the heuristics on how to compile dividends for true/false questions.

The use of certain strategies to solve a mathematical problem indicates the understanding of the relevant concepts. Learners should be able to recognise where and when specific strategies can be used to simplify the processing of a problem [11]. Teachers cannot determine whether a divisibility rule was applied correctly if the correct answer and reason were provided. Participants shift their gaze and attention after practice to more appropriate areas of interest

to perform a relevant task [12]. Therefore, experts inspect more relevant aspects of the stimuli than novices with problem solving [13]. However, analysis of learners' gaze behaviour can be used to determine if they applied the divisibility rule correctly. For example, with dividend 27612 and divisor 6, learners' gaze behaviour could indicate that they inspected only the last two digits (12) of the dividend, motivated that 12 was divisible by 2 and 3, and therefore concluded that 27612 is divisible by 3. Therefore, the gaze behaviour of learners who had the answer and reason correct (A✓R✓) will be used to set the heuristics for the minimum gaze requirements for attention levels per digit per divisor to determine if the learners applied the divisibility rule correctly.

III. EXPERIMENTAL DETAILS

The general pre-post experimental design was used because it requires the same group of individuals to be measured prior to and after the treatment [14]. Learners from two schools in Bloemfontein, South Africa, who had done the divisibility rules, and who, together with their Mathematics teachers, were willing to participate in the project, were selected. School A was located in an upper class urban area, and School B in a township area.

All the learners in Grade 4 to Grade 7 from these schools participated in two paper-based assessments on divisibility rules (Assessments 1a and 2a). Selected learners participated in two eye-tracking assessments on divisibility rules (Assessments 1b and 2b). Revision of the divisibility rules was done between assessments 1 and 2.

A. *Number of participants*

There were 469 learners from School A and 505 learners from School B who participated in both assessment 1a and 2a. There were 78 learners from School A and 77 learners from School B who participated in both assessments 1b and 2b.

B. *Assessments*

Each assessment consisted of fourteen questions - two questions per divisor (divisor 2 to 9, excluding divisor 7). Learners were expected to indicate whether a dividend was divisible by a divisor and provide a reason to motivate their answer. During the eye-tracking assessments, the researcher initiated the move to the next stimulus and recorded learners' verbal responses.

C. *Compilation of dividends*

Learners were presented with dividends consisting of five digits. The dividends were compiled in such a way that the number formed by the last two digits was divisible by the divisor when the dividend was actually not divisible by the divisor, and vice versa. For example, 10199 is not divisible by 9, but if learners inspected the last digit or the number formed by the last two digits, the learners who do not know the divisibility rule would probably indicate that 10199 is divisible by 9. The dividends were compiled in such a way that the sum of the digits varied from 18 to 20.

D. *Research instruments*

A 60 Hz Tobii X2-60W eye-tracker was used to capture the gaze behaviour of learners from School A for assessments 1b and 2b. It was also used for Assessment 1b for learners from School B. Due to a lack of accuracy that was observed during this assessment (possibly due to the fact that the learners all had dark brown irises), a 300 Hz Tobii TX-300 eye-tracker was used for Assessment 2b for learners from School B. The latter eye-tracker handled the limitation observed with the first eye tracker better. The recordings of Assessment 1b for School B were analysed separately and were not compared with Assessment 2b for School B or any of the recordings for School A.

Tobii Studio (version 12) was used to obtain the percentage of total fixation time on each area of interest (AOI) around the digits of the dividend. These percentages, along with learners' responses (answer and reason), were used for statistical analysis. A nine-point calibration was done prior to the assessments to ensure good quality of the gaze recordings.

E. *Stimuli for eye-tracking recordings*

For this study, "Digit 5" refers to the leftmost digit, while "Digit 1" refers to the rightmost digit. The digits were spread out evenly across the display as indicated in Fig. 1. The characters "A" and "B" were added to the front and end of the dividend in order to minimise the positional advantage that end digits might have. Each digit was included in an area of interest (AOI). AOIs were also inserted between the digits and at both ends of the dividend. The fixation time on the AOIs were used to calculate the percentage of total fixation time on the entire dividend for each of the five digits. Fixations on the AOIs between digits contributed 50% to each of the digits on either side.

IV. RESULTS

The average percentage of fixation time per digit for learners from both schools who provided an incorrect reason for the divisibility rule (A×R× or A✓R×) is shown in Fig. 2. The recordings of Assessment 1b for School B were not use because of the lack of accuracy mentioned earlier. The heuristics for compiling a dividend should consider learners' gaze behaviour when they do not know a divisibility rule. When fixations are mainly on the last digits of the dividend, learners' fixations could be correct for divisors 2, 4, 5 and 8, even if they did not know the divisibility rules for these divisors. Teachers must strive to restrict learners from guessing the answer correctly by constructing the dividend in such a way that if a learner does not know the divisibility rule, the learner would probably take a wrong guess.

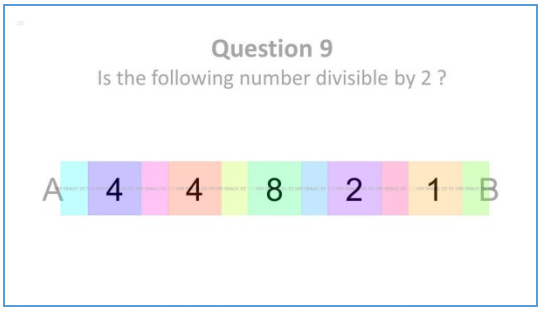


Fig. 1. Example of stimulus used for eye-tracking recordings

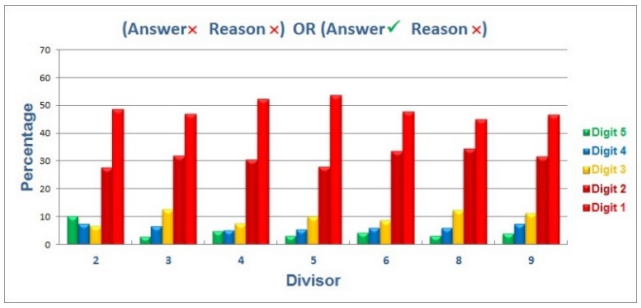


Fig. 2. Percentage of fixation time per digit for learners in A×R× or A✓R×

The following incorrect motivations will be used to compile the heuristics regarding the "do's and don'ts" for dividends that must be divisible by the divisor as summarised in Table I. Table II indicates the heuristics that can be followed to compile the dividend when the dividend must not be divisible by the divisor. Although some of these strategies are the opposite of the strategies stated in Table I, they are listed again.

A. Reasons offered when learners did not know the divisibility rules

The majority of learners from both schools who did not know the divisibility rules, based their reasons for an answer mainly on inspection of the last digit and by stating whether the last digit was divisible by the divisor. Fig. 2 confirms this statement because the fixations of learners who did not know the divisibility rule were mainly on the last digits of the dividend - irrespective of the divisor. These learners devised an incorrect strategy for the divisibility rule and

TABLE I STRATEGIES TO FOLLOW IF THE DIVIDEND MUST BE DIVISIBLE BY THE DIVISOR

Div	Suggested strategies to adopt	Motivation and examples.
2	Let the dividend end in zero (0).	Some learners struggled to identify if zero is divisible by 2. If the dividend ended in zero, learners inspected the last two digits, and the percentage of fixation time will probably be less than the minimum criteria for percentage of fixation time on the last digit. Examples: 31530, 39670, 13590
	Let the dividend start with an odd digit and end in an even digit.	Some learners inspected the first digit of the dividend to determine if it was even. These learners had difficulties to identify an odd or even dividend during the assessments.
	Compile the dividend in such a way that the dividend is even, but there is no digit with value 2.	Some learners inspected the digits of the dividend for the appearance of 2.
	Dividend should not end in 2.	Learners inspected if the last digit is the same as the divisor. Examples: 13138, 93176, 77998
3	Dividend should not end in 3 or a multiple of 3.	Some learners inspected the last digit of the dividend to identify if the last digit was divisible by 3.
	The number formed by the last two digits should not be divisible by 3.	Some learners inspected the number formed by the last two digits to identify if it is divisible by 3.
	The sum of the last three digits should not be divisible by 3.	Some learners added the last three digits of the dividend to determine if the sum is divisible by 3. Examples: 38625, 28317, 43647
4	Dividend should not end in 4 or 8.	Some learners inspected the last digit of the dividend to identify if it was divisible by 4.
	Use a relative high value for the second-to-last digit.	Some learners counted in fours to determine if they reached the number formed by the last two digits. The higher the second-to-last digit, the longer it will take the learner to reach that number. The high value of the second-to-last digit will ensure that learners use mathematical tables to get to an answer.
	The sum of the last two digits should not be divisible by 4.	Some learners added the last two digits to determine if the sum was divisible by 4. Examples: 52796, 37976, 91292
5	Let the dividend end in zero.	Some learners reasoned that the dividend had to end with 5.
	Use a high value for the second-to-last digit if the dividend ends in zero.	Some learners inspected the number formed by the last two digits. If the value of the second-to-last digit is high, learners will spend more time fixating on that digit and probably not meet the minimum criteria for percentage of fixation time on the last digit. Examples: 21690, 74880, 79890
6	Dividend should not end in 6.	Some learners inspected the last digit to determine if it was divisible by 6. Some learners, who knew the divisibility rule, inspected the last digit to determine if it is divisible by 2 and 3.
	The number formed by the last two digits should not be divisible by 6.	Some learners inspected the number formed by the last two digits to determine if it was divisible by 6.
	The sum of the last three digits should not be divisible by 2 and 3.	Some learners calculated the sum of the last three digits and determined if the sum was divisible by 2 and 3.
	Dividend should not end in 23 or 32.	Some learners only inspected the last two digits and reasoned that the digit "3" was divisible by 3 and the digit "2" was divisible by 2.
	The sum of the digits should not be divisible by 6.	Some learners calculated the sum of the digits and determined if the sum was divisible by 6. 73182, 29574, 10404
8	Dividend should not end in 8.	Some learners inspected the last digit to determine if it was divisible by 8.
	The number formed by the last two digits should not be divisible by 8.	Some learners inspected the number formed by the last two digits to determine if it was divisible by 8.
	The sum of the last three digits should not be divisible by 8.	Some learners added the last three digits to determine if it was divisible by 8. Examples: 39160, 79792, 32536
9	Dividend should not end in 9.	Some learners inspected the last digit to determine if it was divisible by 9.
	The number formed by the last two digits should not be divisible by 9.	Some learners inspected the number formed by the last two digits to determine if it was divisible by 9.
	The sum of the last three digits should not be divisible by 9.	Some learners added the last three digits to determine if it was divisible by 9. Examples: 75285, 65817, 34164

stated it with confidence.

For divisors 5 and 8, learners from both schools had difficulty when the dividend ended with zero, for example 68140 / 5 and 23960 / 8. Some indicated that zero was not divisible by 5 or 8, and others based their answer on the number formed by the last two digits.

It was obvious that not all learners knew the mathematical tables as some of them used their fingers when counting. They also made unnecessary errors when doing simple calculations, for example “19 is divisible by 3” and “18 is not divisible by 3”. Besides giving incorrect reasons for specific divisors, some learners in Grade 4 and Grade 5 also exhibited other aspects of general uncertainty or incorrect reasoning:

1) Divisibility depends on whether the dividend is odd or even. The dividends that were used for divisors 4, 6 and 8 were all even numbers because these divisors cannot divide into an odd dividend. If an odd dividend was used with divisors 4, 6 and 8, it would have been correct if learners stated that it was an odd number.
2) Divisibility depends on whether the last digit or last two digits forms a prime number;
3) Tthe last digit should be the same as the divisor for the dividend to be divisible by the specific divisor;
4) The divisor had to be a multiple of the last digit or the last digit had to be a multiple of the divisor to be divisible by the specific divisor;
5) Divisibility depends on whether the divisor was bigger or smaller than the last digit;
6) It is only necessary to inspect the first digit, for example: “the number 44821 is divisible by 2 because 4 is divisible by 2”;
7) The last digit is not divisible by itself, for example: "61254 is not divisible by 4 because the last digit is a 4”;
8) Divisibility depends on whether the divisor is one of the digits of the dividend, for example, “the number 73134 is not divisible by 2 because there is no 2 in the number”;
9) For divisor 3, the sum of the last 2 or the last 3 digits should be divisible by 3;
10) For divisor 4, the sum of the last two digits must be divisible by 4; the last digit must be odd; or the number should end in a 4 or multiple of 4.
11) For divisor 5, the dividend should end with a five to be divisible by 5 (excluding the possibility of ending with 0).
12) For divisor 6, the last digit should be divisible by 2 and 3; the last 2 digits should be divisible by 2 and 3; the last digit should be even/odd; or the sum of the digits should be divisible by 6.
13) For divisor 8, the sum of the last three digits must be divisible by 8;

TABLE II STRATEGIES TO FOLLOW IF THE DIVIDEND MUST NOT BE DIVISIBLE BY THE DIVISOR

Div	Suggested strategies to adopt	Motivation and examples.
2	Let the dividend end in 1.	If the last digit is less than the divisor, learners found it difficult to determine if the last digit is divisible by the divisor. Examples: 79591, 88421, 28261
3	Let the dividend end in 3, 6 or 9.	Learners, who did not know the divisibility rule, inspected the last digit of the dividend to determine if it was divisible by 3.
	The number formed by the last two digits should be divisible by 3.	Some learners inspected the number formed by the last two digits to determine if it was divisible by 3.
	The sum of the last two digits should be divisible by 3.	Some learners added the last two digits to determine if it was divisible by 3.
	The sum of the last three digits should be divisible by 3.	Some learners added the last three digits to determine if it was divisible by 3. Examples: 35933, 17696, 55369
4	Let the dividend end in 4 or 8.	Learners, who did not know the divisibility rule, inspected the last digit of the dividend to determine if it was divisible by 4. Examples: 37574, 48538, 44554
	The sum of the last two digits is divisible by 4.	Some learners added the last two digits to determine if the sum was divisible by 4. Examples: 41722, 95262, 26426
5	Dividend ends in odd number except 5.	Some learners reasoned that the dividend was divisible by 5 because it ended in an odd number. Examples: 31753, 61479, 52417
6	Dividend ends in 6.	Some learners inspected the last digit to determine if it was divisible by the divisor.
	The number formed by the last two digits should be divisible by 6.	Some learners inspected the number formed by the last two digits to determine if it was divisible by 6. Examples: 42766, 13006, 14536
	The sum of the last two digits should be divisible by 3 or 6.	Some learners added the last two digits to determine if it was divisible by 3 or 6. Examples: 32260, 35612, 65348
8	Dividend ends in 8.	Some learners inspected the last digit of the dividend to determine if the last digit was divisible by the divisor. Examples: 82388, 32108, 84348
	The number formed by the last two digits should be divisible by 8.	Some learners inspected the number formed by the last two digits to determine if it was divisible by 8. Examples: 48316, 94124, 16164
	The sum of the last three digits should be divisible by 8.	Some learners added the last three digits to determine if it was divisible by 8. The textbook indicated the divisibility rule for divisor 8 incorrectly by stating that the sum of the last three digits should be divisible by 8. Examples: 88332, 24602, 22116
9	Dividend ends in 9.	Some learners inspected the last digit of the dividend to determine if it was divisible by 9.
	The number formed by the last two digits should be divisible by 9.	Some learners inspected the number formed by the last two digits.
	The sum of the digits should be divisible by 3.	Some learners added all the digits and then determined if the sum was divisible by 3. Examples: 34509, 33999, 60609

14) For divisor 9, the sum of all the digits should be divisible by 3 or the sum of the last 2 or last 3 digits should be divisible by 9.

B. Gaze behaviour of learners who had the answer and reason correct

The learners from School B displayed limited knowledge of the divisibility rules before revision and their improvement after revision was limited as well. Therefore, only the fixations of learners from School A who provided the correct answer and reason (A✓R✓) in Assessment 1b or 2b were used to establish the minimum required percentage time of fixation on the digits of the dividend for each divisor. Table III shows the effect of the digit position on the percentage of fixation time per divisor, as well as the lower limit of the 95% confidence interval. This column was added to determine a trend for the minimum requirements for testing learners' ability to apply divisibility rules.

TABLE III PERCENTAGE OF FIXATION TIME PER DIVISOR AND DIGIT FOR LEARNERS IN A✓R✓ OF SCHOOL A. (LEARNERS SHOULD INSPECT THE UNDERLINED DIGITS ACCORDING TO THE DIVISIBILITY RULE)

			Percentage fixation time	
Divisor	**N**	**Digit**	**%**	**Lower limit of 95% confidence interval**
2	294	A&B	0.32	0.17
		5	2.65	1.81
		4	2.95	2.13
		3	5.91	4.82
		2	33.95	31.16
		1	54.23	51.04
3	105	A&B	0.09	0.02
		5	11.59	9.31
		4	29.45	26.19
		3	22.29	19.88
		2	24.41	21.82
		1	12.17	10.19
4	111	A&B	0.15	0.02
		5	1.90	0.82
		4	2.45	1.40
		3	7.59	5.13
		2	56.94	53.20
		1	30.98	27.22
5	292	A&B	0.28	0.11
		5	2.65	1.74
		4	2.67	1.86
		3	4.70	3.46
		2	39.32	36.31
		1	50.38	47.08
6	84	A&B	0.26	0.03
		5	10.47	8.71
		4	16.60	14.04
		3	18.56	16.42
		2	31.35	28.27
		1	22.76	19.00
8	61	A&B	0.10	-0.04
		5	0.51	0.14
		4	3.02	1.31
		3	32.22	27.37
		2	52.03	47.27
		1	12.11	8.70
9	105	A&B	0.17	0.04
		5	10.28	8.38
		4	15.83	13.96
		3	23.36	20.97
		2	30.11	27.55
		1	20.25	17.61

C. Minimum attention required for divisors 2 and 5

Learners should inspect the last digit of the dividend for both these divisors in order to test for divisibility. Table III indicates that learners spent on average 54.23% for divisor 2 and 50.38% for divisor 5 on the last digit. The lower limits of the confidence intervals indicate that one can be 95% sure that the percentage of fixation time will be higher than 51.04% for divisor 2 and 47.08% for divisor 5. These values were used as the cut-off values for the minimum attention criteria.

The same criteria should be used for divisor 2 and divisor 5 because only the last digit should be focus on. Therefore, learners have to spend 47% of fixation time on the last digit to be 95% sure that the divisibility rule was applied correctly.

D. Minimum attention required for divisor 4

Learners should inspect the last two digits of the dividend in order to test for divisibility. Table III indicates that learners spend 87.92% on the last two digits and the lower limits of the confidence intervals indicate that learners have to spend more than 80.42% on the last two digits. Therefore, the minimum attention required for divisor 4 is that learners have to fixate on both the last and second to last digits, and they had to spend at least 80% of fixation time on these two digits.

E. Minimum attention required for divisor 8

Learners should inspect the last three digits of the dividend in order to test for divisibility. Table III indicates that learners spent 96.36% on the last three digits, and the lower limits of the confidence intervals indicate that learners have to spend more than 83.34% on the last three digits. There were fixations for all the learners (A✓R✓) on digit 2 and digit 3, but only 91% of them fixated on digit 1. Therefore, the minimum attention required for divisor 8 is that learners have to fixate at least 80% on the last three digits and they have to fixate on digit 2 and digit 3.

F. Minimum attention required per digit when learners have to inspect all the digits

Learners should inspect all the digits of the dividend in order to test for divisibility for divisor 3, divisor 6 (if it is even) and divisor 9. Table III indicates that the percentage of fixation time on the digits vary, but there were fixations on all the digits. However, it was found that some learners who had the answer correct and knew how to apply the divisibility rules, only fixated on four of the five digits. Therefore, the minimum attention required for divisor 3, divisor 6 (if it is even) and divisor 9 are that learners have to fixate on at least four of the five digits. Table IV presents a summary of the gaze requirements per divisor.

G. Validation of the minimum required attention levels

The minimum required attention levels were compared with the existing recordings, and 85.74% of responses in A✓R✓ and A✓R× were identified successfully. The false

TABLE IV SUMMARY OF GAZE REQUIREMENTS PER DIVISOR

Div	Requirements
2	The percentage of fixation time on the last digit (digit 1) must be greater or equal than 47%.
3	There must be fixations on at least four of the five digits
4	The total fixation time on the last two digits must be greater or equal than 80%. Learners must fixate on both the last two digits.
5	The percentage of fixation time on the last digit (digit 1) must be greater or equal than 47%.
6	There must be fixations on at least four of the five digits if the dividend is even.
8	The total fixation time on the last three digits (digits 1, 2 and 3) must be higher or equal than 80%. Learners must fixate on digit 2 and digit 3 at least.
9	There must be fixations on at least four of the five digits.

accept rate, where the heuristics would be erroneous if it identified a learner who have applied a divisibility rule correctly, was 8.57%. The false reject rate, where the heuristics would mistakenly indicate that a learner did not apply the divisibility rule correctly, was 5.69% [15].

V. CONCLUSION

When teachers use true/false questions to assess learners on the divisibility rules, there is a high probability that learners can guess the answers correctly. This study uses the rationale behind the learners' reasoning to set heuristics on how to compile dividends per divisor when assessing learners on their ability to apply divisibility rules. This study also set minimum gaze behaviour requirements per digit of the dividend for each divisor as heuristics to determine whether the learners applied the divisibility rules correctly.

These heuristics can be used in a digital classroom where all the computers are equipped with eye-trackers. Questions on divisibility can be displayed on a computer screen and the learners will have to answer by clicking a "Yes" or "No" button. The data will be analysed automatically to determine if the learner has applied the divisibility rule correctly. Teachers can also identify learners who do not apply the divisibility rules correctly. Real-time feedback can be given (i) during introduction of the divisibility rules to identify erroneous strategies before they become habit; (ii) while revision is being done; and (iii) during an assessment to create reports that the teacher can use in a discussion with learners afterwards.

REFERENCES

[1] Singh, U.G. (2014). *The development of a framework for evaluating e-assessment systems* (Doctoral dissertation. University of South Africa). Retrieved from http://uir.unisa.ac.za/handle/10500/14619 (Accessed on 16 April 2016)

[2] Burton, R.F. (2001). Quantifying the effects of chance in multiple choice and true/false tests: question selection and guessing of answers. *Assessment & Evaluation in Higher Education*, *26*(1), 41-50.

[3] Haffejee, F. & Sommerville, T.E. (2014). Fairness in using negative marking for assessing true/false questions. *The Independent Journal of Teaching and Learning*, 9, 75-82.

[4] Obersteiner, A., Moll, G., Beitlich, J.T., Cui, C., Schmidt, M., Khmelivska, T., & Reiss, K. (2014). Expert mathematicians' strategies for comparing the numerical values of fractions - evidence from eye movements. In *Proceedings of the Joint Meeting of PME,* 38, 338-345.

[5] Musser, G.L., Burger, W.F. & Peterson, B.E. (2011). Mathematics for elementary teachers. Ninth Edition. John Wiley & Sons (Asia) Pty Ltd. Asia.

[6] Rożek, B., Błasiak, W., Andrzejewska, M., Godlewska, M., Pęczkowski, P., Rosiek, R., ... & Wcisło, D. (2014). The eye-tracking research method in the process of solving mathematical tasks requiring drawing analysis. *Didactics of Mathematics*, *11*, 43-58.

[7] Zazkis, R., & Campbell, S. (1996). Divisibility and multiplicative structure of natural numbers: Preservice teachers' understanding. *Journal for Research in Mathematics Education 27(5)*, 540-563.

[8] Andrá, C., Lindström, P., Arzarello, F., Holmqvist, K., Robutti, O., & Sabena, C. (2015). Reading Mathematics representations: An eye-tracking study. *International Journal of Science and Mathematics Education*, *13*(2), 237-259.

[9] Storm, C.K., & Zullo, H. (2015). How can Mathematics students learn to play? *Journal of Humanistic Mathematics*, *5*(1), 191-197.

[10] Wall, J., Thompson, C.A., & Morris, B.J. (2015). Confidence judgments and eye fixations reveal adults' fractions knowledge. The annual meeting of the cognitive science society. Retrieved from https://mindmodeling.org/cogsci2015/papers/0441/paper0441.pdf (Accessed on 4 September 2016)

[11] Godau, C., Haider, H., Hansen, S., Schubert, T., Frensch, P.A. & Gaschler, R. (2014). Spontaneously spotting and applying shortcuts in arithmetic - a primary school perspective on expertise. *Frontiers in Psychology*. 5:556. doi 10.3389/fpsyg.2014.00556

[12] Ganor-Stern, D., & Weiss, N. (2016). Tracking practice effects in computation estimation. *Psychological Research*, *80*(3), 434-448.

[13] Jarodzka, H., Scheiter, K., Gerjets, P., & Van Gog, T. (2010). In the eyes of the beholder: How experts and novices interpret dynamic stimuli. *Learning and Instruction*, *20*(2), 146-154.

[14] Duchowski, A.T., (2007). Eye tracking methodology: Theory and practice. Second Edition. Springer. London, England.

[15] Potgieter, P.H. and Blignaut, P.J. 2017. Using Eye-Tracking to Assess the Application of Divisibility Rules when Dividing a Multi-Digit Dividend by a Single Digit Divisor. In Proceedings of SAICSIT '17, Thaba Nchu, South Africa, September 26–28, 2017, 10 pages.

Cloud Robotics Platforms: Review and Comparative Analysis

Viraj Dawarka
Faculty of IT, Design & Communication,
Curtin Mauritius, Telfair,
Moka, Mauritius
vdawarka@curtinmauritius.ac.mu

Girish Bekaroo
School of Science and Technology, Middlesex
University Mauritius, Coastal Road, Uniciti, Flic-en-
Flac, Mauritius g.bekaroo@mdx.ac.mu

***Abstract*—Due to the various advantages that the cloud can offer to robots, there has been the recent emergence of the cloud robotics paradigm. Cloud robotics permits robots to unload computing and storage related tasks into the cloud, and as such, robots can be built with smaller on-board computers. The use of cloud-robotics also allows robots to share knowledge within the community over a dedicated cloud space. In order to build-up robots that benefit from the cloud-robotics paradigm, different cloud-robotics platforms have been released during recent years. This paper critically reviews and compares existing cloud robotic platforms in order to provide recommendations on future use and gaps that still need to be addressed. To achieve this, 8 cloud robotic platforms were investigated. Key findings reveal varying underlying architectures and models adopted by these platforms, in addition to different features offered to end-users.**

Keywords—Cloud-Robotic Platforms, Robotics, Cloud, Knowledge Sharing Platform, Comparative Analysis.

I. INTRODUCTION

Recent advances in cloud computing technologies and robotics have led to the creation of the cloud-robotics paradigm, which relate to networked-connected robots that inherit from parallel computing and data sharing through the Internet. This association is expected to make robots lighter, cheaper and smarter since the utilization of cloud enables heavy processing to be done remotely such that smaller on-board computers are used within robots [1]. For instance, the process for robots to build the map of the environment for localization purposes is both computationally and data intensive and such tasks can be offloaded to the cloud [2]. Concerning data-intensive processes, robots can gain from massive amount of storage space when making use of clouds and also, knowledge acquired by robots can be shared to other such machines [3].

Due to the various benefits of the cloud-robotics paradigm, different platforms have emerged. Cloud-robotic platform enables the build-up of cloud-robotic based robots and some of them are open-source, whilst others are proprietary [4]. Some platforms are composed of a number of high-performance servers, databases, proxy-servers, amongst others [5]. In terms of research, despite the fact that various studies have been conducted regarding the aspects of cloud robotics in many fields, limited published literature is available on the comparison of existing cloud robotic platforms. As related work, one recent study conducted a survey of research on cloud robotics and automation by considering over 150 references in the area [6]. However, this study mentions about cloud robotics platform, limited comparative analysis has been conducted as this was not a major scope the paper. Another study described the development process of cloud robotics as well as the overall architecture of these systems [5]. The current problem that scientists, roboticists and researchers are facing is that an efficient cloud robotic platform is not available to meet the needs of developing an application using existing cloud robotic architectures. Thus, a mixture of platforms ranging from robotic open source software, web services, and cloud platforms are being used to cater the requirements from the chosen architectures which in turn makes it difficult to sustain. Through literatures, cases have been detected where cloud robotic architecture have not been implemented in real life and simulation was used to test the architecture's functionality and efficiency. Since multiple platforms are being used to develop applications using current cloud robotic architectures, the problem of platform dependencies occur and the application fail to function due to one platform not responding to another dependent platform when having flow of data communication. Within this study, cloud robotic platform are highlighted as a key component of the architecture of cloud robotics and although some of the platforms are described, limited comparison has been conducted. Cloud robotic platform in a cloud robotic framework or architecture is the main key component which allows the implementation of the application to further interact and communicate with other components such as robots, cloud or network but going through literature, there is no evidence of which platform is the best fit for this purpose. To address this gap, this paper critically examines and compares existing cloud robotic platforms in order to provide recommendations on future use when implementing cloud robotic architectures and gaps that still need to be addressed. Addressing this gap is expected to provide researchers and experts a comprehensive review of such platforms in addition to recommendations towards improvement of such platforms.

This paper is structured in the successive manner: In the first section, an introduction to the topic is produced then in second section, the methodology used to achieve the purpose is given, followed by a review of existing cloud robotic platform in Section 3. Then, a comparative analysis of the existing cloud-robotic platforms is provided in the fourth section, before making recommendations in Section 5. The work is concluded in Section 6 and avenues for future works are discussed.

II. Methodology

In order to achieve the purpose of this paper, an initial pool of cloud-robotic platforms was searched on Google and Google Scholar by using relevant keywords. The Google search engine was used as it prioritizes results based on relevancy while also employing a variety of techniques to improve search quality through page ranks [7, 8]. The initial search was conducted in March 2018, starting with Google Scholar where 74 conference and journal articles were thoroughly reviewed to assess relevancy, before complementing the search with a general Google search. Keywords such as Cloud robotic platform, cloud platform, robotic platform were used to search the database then the exact name of the cloud robotic platform was utilized to further narrow down the search. Finally, only 8 cloud-robotic platforms were identified, whilst the other sources were principally related to frameworks and state-of-the-art review of cloud-robotics. Once the platforms were identified, literature search pertaining to each platform was thoroughly conducted by going through relevant articles while also finding information on key websites. The information gathered was then analyzed and is presented in the next sections of this paper. Similar methodology was used in different studies performing comparative analysis [9, 10]. Using this methodology enables to review all the research work pertaining to this topic in the literature to find an accurate answer to the research problem.

III. Review of Cloud Robotic Platforms

Using the methodology defined in the previous section, different cloud-robotic platforms were identified. These are discussed as follows:

A. Rapyuta

Also known as the RoboEarth Cloud Engine, Rapyuta is an open-source cloud robotics platform [9]. It is based on an elastic computing model and active distributed secure computing environments in which robots are deeply associated, while allowing robots to contribute most of their services to other robots [10]. This platform is known to eliminate complexity, costs, possibility of deploying, interfacing and managing robotics systems so that more time is available to do other tasks [9]. By allowing access to the RobotEarth Knowledge Repository, Rapyuta allows robots to store and share information, offload computation and collaborate and achieve a common task [10]. Moreover, a range of capabilities is provided such as disk quota, I/O limits and memory limits configuration, among others. Furthermore, it also allows the outsourcing of around more than 3000 Robot Operating System (ROS) packages and is extensible to other robotic middleware [11]. Additionally, a recent work using this platform involves the pre-installation of Amazon Machine Image (AMI) which can launch the Rapyuta in one of the Amazon's data center in a short period of time and permits the robots to authenticate themselves, create one or more environments in the cloud and launch the process [11].

B. Robot Operating System (ROS)

The ROS platform enables the production of software modules in order to execute typical robot activities such as object recognition [12]. The concept of "not re-inventing the wheel" is the principal aim of the ROS platform, where it provides integrated libraries that are easy to use in addition to multitude facilities such as manipulation, navigation control, and hardware abstraction for sensors and actuators, among others [13]. ROS also gives the advantages of inter-platform operability between multiple programming languages such as C, Java and Python. As part of this platform, ROS processes or Nodes involve data processing in the platform and a message-passing distributed system derived on the publish/subscription paradigm is achieved by ROS where Nodes produce messages on Topics which other Nodes employ [14]. In terms of application, this platform has provided solution for real-time ball trajectory tracking for tennis and football events through the creation of an environment with the integration of Open Source Computer Vision libraries (OpenCV) for object detection and tracking [15]. At present, a newer version of ROS, namely ROS 2, is being actively developed to improve cross-platform support capability [16].

C. C2RO Cloud Robotics

Established in 2016 in Montreal, C2RO Cloud Robotics is a cloud-based software robotics platform for the global service robotics target market [17]. The C2RO platform connects robots using patent-pending technologies and augments the capabilities of robots through a fast and secure communication. It also provides a robot-agnostic software-as-a-service (SaaS) platform that utilizes an information processing technology, which functions in a real-time manner to grant robots an artificial intelligence solution in a secure, fast and inexpensive approach [18]. This platform was created in order to address the industrial automation demand comprising of problems such as lack of robots' connectivity that result in the inability for monitoring real time problem, the limitation of pre-programmed tasks due to limited onboard sensing and computing power as well as non-cooperation of robots [18]. In order to address these problems, the C2RO platform upgrades processing power via a hybrid solution of completing high-skilled tasks and a cloud-based robotics platform where multiple robots can share knowledge instantly across multiple sites and geographies [17].

D. Microsoft Robotics Developer Studio (MRDS)

Released by Microsoft in 2006, the MRDS platform enables programming robots in the Windows environment and it can interact with the circuits commonly known as microcontrollers on the robots to control actuators over a hardwired link or Bluetooth [19]. MRDS contains a .NET-based service-oriented runtime, comprising of components such as Concurrency and Coordination Runtime (CCR) and Decentralised software services (DSS) [20]. This platform concentrates in making robotics applications that allow either to be simulated by using Visual Simulated Environment (VSE) which is 3-D virtual simulator or through Visual Programming Language Environment (VPL), a programming interface that connects to real robots. MRDS is considered as a crucial product when bestowing Service-oriented architecture (SOA) framework for Robot as a Service (RaaS) in cloud computing to embedded systems [21].

E. REALabs

REALabs is a cloud robotic platform that enables the computer running the robotic application and mobile robot to interact with each other over the network [22]. This platform was

built based on the Platform as a Service (PaaS) model [23]. REALabs is completely based on Web technologies and consists of four main software packages from its architecture, namely, the front-end package, the protocol handler, embedded package and the management package. The platform has been mainly used in Web Labs over public internet where the user implements the robotic application on personal computer so as to control the robot through the network [22]. Several updates have been conducted on the platform such as integration of Web services in the end of 1990s, HTTP/XML-based Remote Procedure Call and transition to the Representational State Transfer (REST) architecture [23].

F. Rospeex

Rospeex is a cloud robotic platform for multilingual spoken dialogues with robots for ROS developed by the National Institute of Information and Communications Technology (NICT) [24]. It is equipped with a straightforward interface for speech synthesis and speech recognition in different languages including Japanese, English [25]. This platform is free for use by roboticists and does not require authentication. Moreover, the platform comes with a bundle including a browser user interface, the Rospeex cloud services and the Rospeex modules comprising of voice activity detection, noise reduction and speech synthesis. The user interface of the rospeex platform has been developed in HTML5 and can operate on many platforms such as Linux, Windows and Android smart devices. Also, two types of users can adopt this platform, namely user and developer and with the use of Rospeex, large amount of robotics-related statements can be stored on the cloud server [24].

G. DAvinCi

DAvinCi stands for Distributed Agents with Collective Intelligence and was built in Singapore by the ASORO laboratory to produce 3D-models of environments for robots during simultaneous localization and mapping (SLAM) [21]. It supports an augmented architecture for large environment to allow group of robots to operate in large environments [26]. DAvinCi consists of three technologies, namely, the open source Hadoop Distributed File System (HDFS), ROS architecture and the Hadoop Map/Reduce Framework. The ROS architecture is used for sensor data collection and to communicate between clients and robots, the HDFS is used for data storage and Hadoop Map/ Reduce framework is used for batch processing of visual information and sensor data. An established way of communication and messaging between the robots and the DAvinCi server is provided by the ROS and its goal is to offload huge data workloads from the robots to a backend cluster system [26].

H. GostaiNet

GostaiNet was developed by the French robotic firm called Gostai and enables robots to perform face detection, speech recognition and other task remotely [27]. Seamless control of any robot from anywhere around the world is provided through a web browser and services are hosted by Gostai on the GostaiNet robotics cloud [28]. This architecture gives the opportunity to decentralize artificial intelligence in order to produce economic robots with complex behaviors and with autonomous capabilities [29]. The latest work that was made available is the Jazz robots that were implemented on Gostai's Urbi open-source ROS while also using the GostaiNet cloud-computing infrastructure to make cloud based video recording [30].

IV. Comparative Analysis

Review of the cloud-robotic platforms showed that most of them are open-source besides REALabs and DAvinCi platforms. While being open-source, different such platforms evolve continually as developers and other contributors keep on adding or updating features. This also promotes adoption of such solutions by developers, who can also obtain support from the community. Use of such platforms are also promoted by the fact that most platforms provide a user-friendly user interface where user can offload robotic data onto the cloud service. On the other hand, while being open-source, some platforms such as ROS do not have up-to-date documentation for its users due to associated maintenance costs. Similarly, most platforms enable sharing to other robots through the cloud environment, besides MRDS. When sharing data or knowledge to robots built in a different platform, a key challenge faced involves format for representing and exchanging data [6]. For instance, although sensor data in the form of images have a small number of popular formats, trajectory-related data have no standardized format [31, 32]. Moreover, Rapyuta and ROS were found to be highly compatible with other platforms as discussed earlier, whilst REALabs, showed to have the lowest compatibility. In TABLE I, Compatibility with other platforms has been discussed in terms of cross platform capabilities where a scale high means the platform can be used by a large number of other platforms, moderate means the platform can be utilized by an average number of platforms and low for a very minimal number of platforms. The comparative summary of the review conducted is given in TABLE I.

TABLE I. Comparative analysis

Cloud Robotic Platform	Security	Open Source	Underlying Model or Architecture	Compatibility with other platforms	Sharing to other robots
Rapyuta	Provides a secure customizable computing environment in the cloud to offload heavy computation.	Yes	Elastic computing model	High	Yes
ROS	A cryptographic method has been implemented in [35] to secure the ROS communication channels.	Yes	Publish/subscribe message passing architecture	High	Yes
C2RO	Virtual Barriers is used for data access control, Secure	No	Hybrid cloud robotics model	Moderate	Yes

	Sockets Layer and Transport Layer Security are used to avoid eavesdropping and "man in the middle" attacks.				
MRDS	The MRDS studio has a security manager page which manages authentication for users.	No	Service Oriented Architecture	Moderate	No
REALabs	A secure Single Sign On service based on SAML (Security Assertion Markup Language) for user authentication is provided [36].	Yes	Platform as a Service (PaaS) model	Low	Yes
Rospeex	A ROS node is implemented for security and network reasons so as not to conduct noise reduction and Voice Activity Detection.	No	Node structure model	Moderate	Yes
DAvinCi	A fail safe mechanisms is being worked on the communication between the DavinCi server and the robots during transfer of messages.	Yes	Combination of distributed ROS architecture, the open source Hadoop Distributed File System(HDFS) and the Hadoop Map/Reduce Framework	Moderate	Yes
GostaiNet	Image analysis within the system allows movement detection for the robots making it easy to detect intruders.	Yes	GostaiNet cloud computing architecture	Moderate	Yes

Furthermore as shown in Table I, different such platforms have varying architectures or models being utilized, while also having dissimilarity on focus of the platforms in terms of characteristics and abilities. For instance, MRDS is the only platform which provides two powerful engines for graphics and dynamics in order to allow the development of distributed and concurrent processes in an innovative way [20]. This fundamentally provides the advantage of developing service-oriented, concurrent and asynchronous applications with a myriad of programming languages, including the visual programming language. REALabs platform is based on the PaaS model and uses REST, which is an alternative to web services in order to allow interaction between virtual machines and mobile robots [22].

While operating PaaS, REALabs is deployed on a Virtual Machine where latency becomes a concern. Together with this issue, the mobile robots situated on the field do not allow the usage of cloud services in a variety of mobile robotics applications available [22]. As for ROS, it is based on the Publish/subscribe message passing architecture and the platform ensures a clean programming standards to the user by allowing threads in the application to publish and subscribe to messages only. ROS provides modularity when implementing robot applications where in case a component (e.g. a sensor or a motor) crashes, the entire application does not crash due to connection with a distributed message system. DAvinCi, as discussed earlier, uses a combination of architectures in order to provide improved performance of the system through the incorporation of several computing nodes. However, the main concern is that if one of the component does not work (e.g. DAvinCi server), problems may arise if improper backup mechanisms are unreliable. Another observed issue is that performance of the overall system varies based on the number of nodes inculcated in the system. Rapyuta, in turn, is based on the elastic computing model and provides a friendly interface when offloading robotic data while also providing access to a repository of shared knowledge amongst robots. Furthermore, a proper security aspect is imparted to each robot in this platform and provides a bidirectional communication with the robots. Finally, Rospeex is based on the node structure model and is dedicated only to speech capabilities of robots using browser user interface. Its modules are related to voice activity, speech synthesis and noise reduction together with cloud services. Since Rospeex uses its own cloud service, it becomes a disadvantage when it comes to multi-platform operation. Moreover, it was also highlighted in a previous study that the waiting time for speech recognition processing time in Rospeex is not appropriate as it can deteriorate the dialogues' interactive aspect [26].

V. Recommendations

The comparative analysis performed showed different limitations of existing cloud-robotic platforms and for each platform, recommendations are proposed as in TABLE II, for further research and improvement by research community.

TABLE II. RECOMMENDATIONS

Cloud Robotic Platform	Reported Issue	Recommendation
Rapyuta	High Computational latency	Colocation data centers can be integrated with the platform to ensure an exceptional network coverage.
ROS	Since ROS is an open-source framework, it is vulnerable to authentication, authorization and insecure communication issues. Lack of up-to-date documentation.	A cryptography algorithm such as Diffie-Hellman or Message Authentication Code (MAC) could be introduced in the framework for strengthen the security aspect of the framework. To provide up-to-date documentation for end-users.
C2RO	A large computational power is required for the robots when using visual simultaneous localization and mapping (SLAM) to locate themselves.	An Artificial Intelligence (AI) application can be implemented to increase the computation power of the robot so as it can be more autonomous when taking important decisions regarding localization and mapping.
MRDS	Incomplete models when simulating real time situation such as wheeled vehicles and modelling sonar.	More research could have been done in this domain where an accurate model can be proposed using MRDS platform to strengthen the capacity of real time processing in daily life.
REALabs	REALabs is based on RESTful interfaces, meaning it relies on HTTP for its security aspect making it vulnerable.	A more secure protocol can be used in the framework such as OAuth1.0a to provide secure web services using the RESTful interfaces.
Rospeex	Waiting time for speech recognition processing time is not appropriate Uses its own cloud service, which is a disadvantage when it comes to multi-platform operation.	Segmenting the Rospeex module that sends the speech file to the server into multiple fragments and sends them one by one for processing. Could integrate the use of several form of cloud service where it will be easier for a myriad of platform to operate when implementing applications.
DAvinCi	Owing to the utilization of Hadoop map reduce computing cluster, a high bandwidth usage is required.	A hosted filtering can be integrated in the framework between the DaVinCi server and public cloud to filter spam messages that are transmitted over HTTP.
GostaiNet	The provided services are proprietary as they are hosted by Gostai on the GostaiNet robotics cloud.	The services can be deployed in collaboration with a public cloud service to make it more accessible to many applications.

VI. Concluding Remarks

This paper examined and compared 8 cloud robotic platforms used by roboticists around the world when developing cloud robot applications. The review showed that most of the platforms are open-source, thereby inheriting some advantages and disadvantages of being open-source. Moreover, some platforms such as Rapyuta and ROS were found to be highly compatible with other cloud-robotic platforms while enabling integration of features. In addition, different such platforms were found to have varying underlying architectures or models, while also having dissimilarity on focus in terms of characteristics and abilities. Depending on the framework or architecture that will be used, each platform will act differently and will try to blend accordingly to give an efficient product. As future work, the same platforms will be further analyzed following application to develop cloud robotic applications. This will also help to increase the number of comparison criteria studied in the comparative analysis and an efficient platform can be provided for the implementation of cloud robotic application using the prescribed recommendation.

References

[1] E. Guizzo, "Robots with Their Heads in the Clouds," IEEE Spectrum, 2011. [Online]. Available: http://spectrum.ieee.org/robotics/humanoids/robots-with-their-heads-in-the-clouds. [Accessed 5 Mar 2018].

[2] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: part I," IEEE robotics & automation magazine, vol. 13, no. 2, pp. 99-110, 2006.

[3] G. Hu, W. Tay and Y. Wen, "Cloud robotics: architecture, challenges and applications," IEEE network, vol. 23, no. 6, 2012.

[4] K. Sugiura and K. Zettsu, "Rospeex: A cloud robotics platform for human-robot spoken dialogues," in 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2015.

[5] J. Wan, S. Tang, H. Yan, D. Li, S. Wang and A. Vasilakos, "Cloud robotics: Current status and open issues," IEEE Access, vol. 4, pp. 2797-2807, 2016.

[6] B. Kehoe, S. Patil, P. Abbeel and K. Goldberg, "A survey of research on cloud robotics and automation," IEEE Transactions on automation science and engineering, vol. 12, no. 2, pp. 398-409, 2015.

[7] S. Grin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," Computer networks and ISDN systems, vol. 30, no. 1-7, pp. 107-117, 1998.

[8] K. K. Lavania, S. Jain, M. K. Gupta and N. Sharma, "Google: A Case Study (Web Searching and Crawling)," International Journal of Computer Theory and Engineering, vol. 5, no. 2, pp. 337-340, 2013.

[9] A. Zakari, A. Lawan and G. Bekaroo, "Towards Improving the Security of Low-Interaction Honeypots: Insights from a Comparative Analysis," in International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering, 2016.

[10] B. Kim and R. Neff, "Measurement and communication of greenhouse gas emissions from US food consumption via carbon calculators," Ecological Economics, vol. 69, no. 1, pp. 186-196, 2009.

[11] Rapyuta, "A cloud Robotics Platform," Rapyuta, 2018. [Online]. Available: http://www.rapyuta.org/. [Accessed 7 Apr 2018].

[12] D. Hunziker, M. Gajamohan, M. Waibel and R. D'Andrea, "Rapyuta: The roboearth cloud engine," in 2013 IEEE International Conference on Robotics and Automation (ICRA);, 2013.

[13] G. Mohanarajah, D. Hunziker, R. D'Andrea and M. Waibel, "Rapyuta: A cloud robotics platform," IEEE Transactions on Automation Science and Engineering, vol. 12, no. 2, pp. 481-493, 2015.

[14] L. Garber, "Robot OS: A new day for robot design," Computer, vol. 46, no. 12, pp. 16-20, 2013.

[15] S. Chitta, "MoveIt: an introduction," in Robot Operating System (ROS), 2016.

[16] ABI Research, "Robot Operating System Evolving to Foundational Software Platform," ABI Research, 2014. [Online]. Available: https://www.abiresearch.com/press/robot-operating-system-evolving-to-foundational-so/. [Accessed 2 May 2018].

[17] H. Yadav, S. Srivastava, P. Mukherjee and B. Lall, "A real-time ball trajectory follower using Robot Operating System," in IEEE 2015 Third International Conference on Image Information Processing (ICIIP), 2015.

[18] D. Hood and W. Woodall, "ROS 2 update – summary of alpha releases, architectural overview," in ROSCon 2016, Seoul, Korea, 2016.

[19] J. Park, "What is Cloud Robotics?," C2RO Cloud Robotics, 2018. [Online]. Available: http://c2ro.com/what-is-cloud-robotics/. [Accessed 7 May 2018].

[20] Robotics Business Review, “C2RO Cloud Robotics,” Robotics Business Review; 2018. [Online]. Available: https://www.roboticsbusinessreview.com/listing/c2ro-robotics/. [Accessed 7 May 2018].
[21] K. Johns and T. Taylor, Professional microsoft robotics developer studio, John Wiley & Sons, 2009.
[22] J. Jackson, “Microsoft robotics studio: A technical introduction,” IEEE robotics & automation magazine, vol. 14, no. 4, 2007.
[23] A. Koubaa and E. Shakshuki, Robots and Sensor Clouds, Springer International Publishing, 2016.
[24] A. do Nascimento, E. Cardozo, R. Souza and E. Guimarães, “A platform for cloud robotics,” IFAC-PapersOnLine, vol. 49, no. 30, pp. 48-53, 2016.
[25] E. Guizzo and T. Deyle, “Robotics trends for 2012,” IEEE Robotics & Automation Magazine, vol. 19, no. 1, pp. 119-123, 2012.
[26] K. Sugiura and K. Zettsu, “Rospeex: A cloud robotics platform for human-robot spoken dialogues,” in 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2015.
[27] Rospeex, “What is rospeex?” Rospeex, 2018. [Online]. Available: http://rospeex.org/top/. [Accessed 6 May 2018].
[28] R. Arumugam, V. Enti, L. Bingbing, W. Xiaojun, K. Baskaran, F. Kong, A. Kumar, K. Meng and G. Kit, “DAvinCi: A cloud computing framework for service robots,” in 2010 IEEE International Conference on Robotics and Automation (ICRA), 2010.
[29] Sai Natkar (“Cloud Robotics: It’s time to offload their brain on Cloud, for better Robotic future!” ESDS, 2018. [Online]. Available: https://www.esds.co.in/blog/cloud-robotics-its-time-to-offload-their-brain-on-cloud-for-better-robotic-future/#sthash.lvdbLgds.dpbs. [Accessed 10 May 2018].
[30] G. Mester, “Cloud Robotics Model,” Interdisciplinary Description of Complex Systems: INDECS, vol. 13, no. 1, pp. 1-8, 2015.
[31] R. Heuillard, “Gostai: la robotique et la téléprésence abordables,” Clubic Tech, 2009. [Online]. Available: https://www.clubic.com/actualite-251396-gostainet-robotique-personnelle-telepresence.html. [Accessed 7 May 2018].
[32] H. Lichtman, “Gostai Jazz Telepresence Robot Unveiled,” Telepresence Options, 2010. [Online]. Available: http://www.telepresenceoptions.com/2010/12/gostai_jazz_telepresence_robot/. [Accessed 7 May 2018].
[33] E. Prestes, J. Carbonera, S. Fiorini, V. Jorge, M. Abel, R. Madhavan, A. Locoro, P. Goncalves, M. Barreto, M. Habib and A. Chibani, “Towards a core ontology for robotics and automation,” Robotics and Autonomous Systems, vol. 61, no. 11, pp. 1193-1204, 2013.
[34] M. Tenorth, A. Perzylo, R. Lafrenz and M. Beetz, “Representation and exchange of knowledge about actions, objects, and environments in the roboearth framework,” IEEE Transactions on Automation Science and Engineering, vol. 10, no. 3, pp. 643-651, 2013.
[35] B. D. S. Benjamin Breiling, "Secure Communication for the Robot Operating System," in 11th Annual IEEE International Systems Conference, Montreal, April 2017.
[36] E. G. L. R. R. S. F. P. a. F. P. Eleri Cardozo, "A Platform for Networked Robotics," in IEEE/RSJ International Conference on Intelligent Robots and Systems , Taipei, Taiwan, October 18-22, 2010.

Web Authentication Security Using Image Steganography and AES Encryption

Hope Mogale*, Michael Esiefarienrhe† and Lucia Letlonkane‡
Department of Computer Science, North-West University
Private Bag X2046, Mafikeng 2735, South Africa
Email: *Hope.Mogale@ieee.org, †Michael.Esiefarienrhe@nwu.ac.za, ‡Lucia.Letlonkane@nwu.ac.za

***Abstract*—In this paper we present Web based highly secure Authentication Security system that uses Image Steganography and 128-bit Advanced Encryption Standard algorithm. Our developed web authentication system encrypts users passwords with an AES encryption that takes octillion years to decrypt and is coupled with face identification photograph which is used as a stegoimage to hide the encrypted password for more robust security. When put to test our Web Authentication Security systems shows positive results which outstand advanced steganalysis attacks such as the chi-squared attack and neighbourhood histogram. We recommend this security for prospective web applications of the future that will be handling sensitive user information.**

Keywords - ***AES Encryption, Steganography, Web, Authentication, Cryptography***

I. Introduction

Over the years the internet has grown to become an enormous technological infrastructure for modern cloud applications. Many companies have now migrated from deploying software and selling it in hard copies to selling it and hosting over the internet. But how safe is the internet? And can it be trusted with intellectual property worth billions of dollars? Most importantly can it be trusted with sensitive user information. The Internet and the World Wide Web (W3) in its bare form carries a major risk without security. Without proper security maintained, sending data and all kinds of information over the Internet can be a big risk and result in data loss, corruption or theft. The world is experiencing a high rise in cybercrime and therefore security on data needs to be always up to date because data requires protection and proper security for sending and receiving over the Internet. Data or information needs to be first concealed with proper security before sending it on the Internet or any communication mediums that utilizes the internet. To alleviate this problem we can use Steganography which is the art and science that provides us with the methods of embedding messages sent over communication mediums in digital media such as digital images, audio, and video for more robust security. Steganography is the art of hiding information in information. This can also be referred to as hiding information in plain sight. Steganography hides the fact that information or communication exists [1] in the first place and this helps prevent hackers from compromising information or communication since they can't prove it exists. Steganography is different from cryptography in that, cryptography provides methods of encryption to protect data from being compromised while steganography eliminates the fact that the data exists and this makes steganography even more secure. Steganography and cryptography can both be combined and one can employ the other for the case when the data is compromised and more security is needed to protect the data. There are many different forms of steganography as illustrated in figure 1, there is image steganography, audio steganography, text-steganography, and video steganography. Steganography uses media that is used to hide information in plain sight and these are usually called covers [2]. The most widely used covers are image covers because of their consistency and how easy it is to obtain one and also to transmit over the internet. Another reason is because of the compression standards that image covers often can be found on many formats. However, in our research we focus on JPEGs mainly because of their popularity over the internet but most importantly because their compression standard which is very low and makes embedding simple and none suspicious because the final stegoimage is usually small in size. JPEGs allow secret messages to be embedded in the Least Significant Bits (LSB) of the image and hidden inside images and be sent over the Internet with no interference at all as seen in figure 2 below. The message can be encrypted for the case when the communication is tampered or even eavesdropped and this the method we shall adopt on our research. Solid steganography relies on good algorithms [3].

A. Least Significant Bit Steganography

Least Significant Bit Steganography seen in figure 2 is a method of steganography that provides a simple way to embed messages in different covers. LSB method is common in audio and image covers because of their robustness and the wide range of compression standards available to them. LSB method involves manipulating bits of a cover and configuring some to the bits of the cover file to accommodate the bits which contribute the overall message [4]. However LSB method can be very complex and lead to undesired results depending on the nature of the cover and the approach undertaken when implementing it. If the cover is too large then this might compromise the security of the steganography and may subject it steganalysis attacks. To overcome this shortcoming embedded covers can be chosen based on metrics such as size and type of image used. The quality of the image also plays a huge role in LSB Steganography because if the image cover

embedding is low in quality it will compromise the security of the message and can be subjected to steganalysis attacks. We use encryption for a more robust security on our payloads.

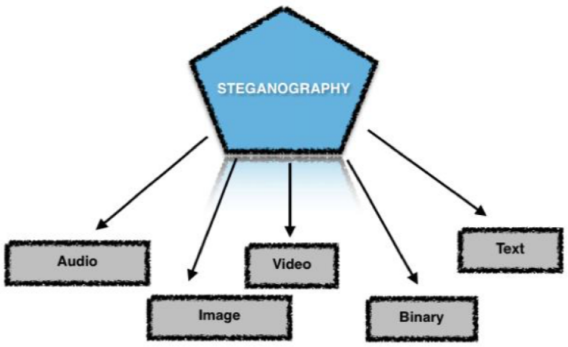


Figure 1. Types of Steganography

LSB method is quite common in fragile Steganography, and different approaches exist for implementing LSB [5] and this is normally decided by the cover chosen to be the final message carrier. There almost infinite number of digital images of various types that can be found on the Internet and these images can either be black and white, Grey or full RGB colour. However, for our method we have chosen to only utilize full RGB colour JPG images because this allows us to use the two dimensional Discrete Cosine Transform (DCT) for each colour component of the JPEG image. Since pixels of a JPEG image can be represented using a two dimensional matrix. We can assume that this matrix has finite rectangular support on $[0, N_1, -1]x[0, N_2 - 1]$ and a two dimensional DCT as:

$$X_C(k_1,k_2)\triangleq\sum_{n_1=0}^{N_1-1}\sum_{n_2=0}^{N_2-1}4x(n_1,n_2)\cos\frac{\pi k_1}{2N_1}(2n_1+1)\cos\frac{\pi k_2}{2N_2}(2n_2+1),$$
$$for\quad (k_1,k_2)\in[0,N_1-1]x[0,N_2-1];Otherwise,X_C(k_1,k_2)\triangleq 0. \qquad (1.1)$$

We can further use DCT to transform successive 8 x 8 pixel blocks of the image into 64 DCT coefficients for which later we can quantize the coefficients for more robust security. High frequencies usually occur at the highest indices $(k1, k2)$ of (1.1). The important thing to note is that the inverse for the DCT is given as $(n_1, n_2) \in [0, N_1, -1]x[0, N_2 - 1]$ also exists and this works to our advantage because we can rebuild stegoimages without experiencing redundancies.

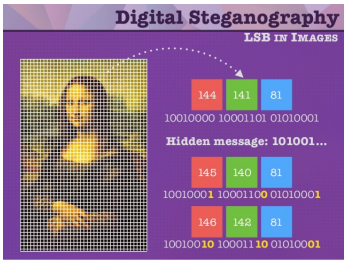


Figure 2. Illustration of LSB Steganography

We can also employ both Peak Noise to Signal Ration and the Mean Squared Error (MSE) to discriminate between the original image before embedding and the final image after embedding in the least significant bits. If we assume an image of size $B * C$ then MSE can be described as follows:

$$MSE = \frac{1}{B\,X\,C}\sum_{i=1}^{B}\sum_{j=1}^{C}[I(i,j) - I\hat{}(i,j)]^2 \qquad \textbf{(1.2)}$$

Steganography can also be cover based and not cover based as seen in figure 3. However, we will only focus on cover based steganography that utilizes JPEGs digital images as covers.

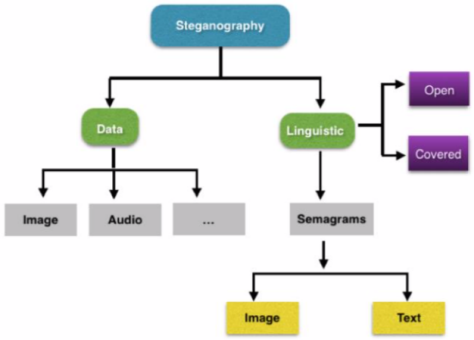


Figure 3. Stratums of Steganography

Each digital image has a structure that is composed of picture elements (pixels). Black and white image pixels have a single bit that can either be zero or one. For grey scale, images each pixel is defined in 8bits of 0 (black pixels) to 255 (white pixels) and for Full colour each pixel has 24bits each of 8bits emphasizing RGB. To embed on these pixels generally we can simply use:

$$qi = 2\left|\frac{ki}{2}\right| + mi \qquad \textit{(1.3)}$$

Where qi, ki, and mi represent the $i-th$ bit of the message and the $i-th$ pixel before embedding on the image cover. LSB method can also be applied on audio signals [6] but we will delve more into this on our future work. LSB has infinite possibilities and can be both detrimental and beneficial as a method for security on information.

B. AES Cryptography

Cryptography is the science of concealing information using codes. Cryptography enables secure communication in the presence of third parties since most of the communication will be encrypted and unreadable to the third parties. There are different types of cryptographies and these differ in that

one uses symmetric keys, asymmetric keys or hash functions. There are different methods of cryptography dependent on these keys and these include DES, AES, RC5, ISA, SHA-1, RSA and etc. For our research we will only focus on employing the Advanced Encryption Standard (AES).

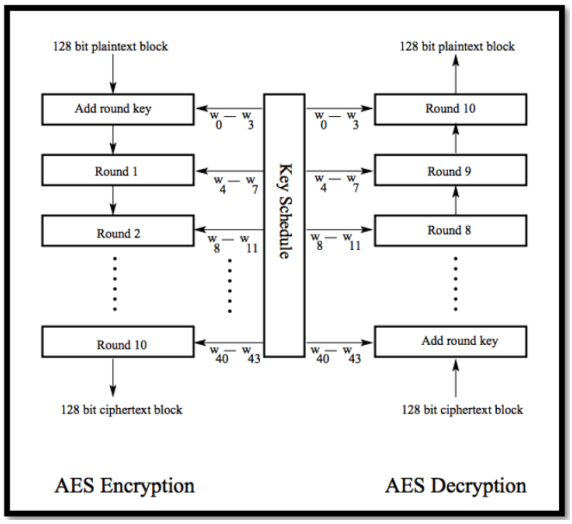


Figure 4. Illustration of AES Encryption

AES Encryption uses a block cipher with block length of 128 bits as illustrated in figure 4. AES uses the notion of a word and since word is four bytes, the overall length would be 32 bits. Hence, each column of the state array is a word, as is each row. For 128 bit keys encryption usually consists of 10 rounds of processing and this increases as the number of keys increase. AES algorithm consists of three stages which is the initial round when the encryption is initiated and at this stage each byte of the state array is combined with a block of the round key using a bitwise XOR. Then follows the next round which executes the $SubBytes$, $ShiftRows$, $MixColumns$, and $AddRoundKey$ subroutine. Lastly follows the final round which executes the $SubBytes$, $ShiftRows$, and $AddRoundKey$ subroutines. According to the United States Government the strength of all key lengths of the AES algorithm are sufficient for protecting classified information. Hence, why we will employ AES for encryption on web authentication security.

II. Challenges of Web Authentication Systems

There are many identified challenges that can be experienced by a web authentication system ones we choose to alleviate and overcome are ones we list below:

1) **Explicit Passwords** – Users are often required to enter passwords before being authenticated even though this is often masked this is not sufficient enough as it does carry risk in that the perpetrator can identify and guess the length of the password and often develop a method to compromise this information such as brute force attacks. We intend to solve this by using AES encryption.
2) **TLS/SSL Computation Complexity** – In most web authentication systems secure communication is usually established by use of TLS/SSL encryption however this requires the whole communication be encrypted. Since we are concerned only with authentication this will not be necessary.
3) **SQL Injection** – Most web authentication systems experience an attack called SQL injection in which the attacker compromises application code to access or corrupt database content. We are using steganography as a means to dismiss the idea of an existing database in which the database can be compromised. Even if it is compromised steganography will protect the information and hence this will make it hard for attackers to be able to steal users information.
4) **HTTP authentication Headers** – Web authentication applications usually utilize HTTP authentication headers to convey a request for users credentials. The main problem is when these headers are not secure enough or not expressive enough to gain users trust. We provide a solution to this by requesting the users photograph as a means of confirmation for identity.

III. Related Work

There has been substantial amount of research that has been conducted by researchers and scholars across the world in the past and recently. The ones we deem fit for our needs are the ones we discuss below.

Image steganography can be highly secured or easily compromised using steganalysis hence why Sakar et. al in [7] proposed a novel technique for image steganography based on Block-DCT and Huffman Encoding. In their work DCT is used to transform original image (cover image) blocks from spatial domain to frequency domain for more security. Then the Huffman encoding is also performed on the secret messages/images before embedding and each bit of Huffman code of secret message/image is embedded in the frequency domain by altering the least significant bit of each of the DCT coefficients of cover image blocks. The results from their experiment show The experimental results show that the algorithm has a high capacity and a good invisibility.
Moreover PSNR of cover image with stegoimage shows the better results in comparison with other existing steganography approaches. We find this work amazing and aim utilize similar approach with DCT. Al- Shatnawi [8] proposed a new method for hiding a secret message based on searching and finding identical bits between the secret messages and image pixel values. Their method was able to achieve 83% accuracy ration while remaining efficient, simple, fast, robust to attacks and help improve the overall image quality. Yusik et. al in [9] proposed Web based image authentication using invisible Fragile watermark which is a web based image authentication

method based in digital watermarking. Jain et. al [10] proposed a method of using edges of the images for hiding text messages in image Steganography. Their research demonstrates how the edges of the images can be used for hiding text message in Steganography. Their work also gives the depth view of image steganography and Edge detection Filter techniques.

In [11] researchers proposed an efficient steganalytic LSB matching based on image noise. The obtained experimental results demonstrates that on low embedding ratios the detection accuracy can be high. This also tells us that embedding payloads on image noise is not ideal and this is supported by Zhang at. el in [12] who proposes a novel universal steganalyzer for additive noise steganography in JPEG decompressed images. Hiding information on image noise is not secure enough and may be subjected to defeat with steganalysis attacks. An Analysis of LSB and Discrete Cosine Transform (DCT) has been conducted by Walia and Jain in [13] and the analysis show that DCT out performs LSB in terms of high ratios when evaluating using PSNR. We choose to ignore this since our approach is based on the spatial domain and most importantly on the web domain. Hossain et. al in [14] proposed perhaps the most important work which shall play a crucial part in our work. They proposed an efficient filtering based approach for improving LSB Image steganography using Status Bit along with AES Cryptography. For improving their security they have chosen to use an AES cryptography technique before applying Steganography.

AES cryptography is used to change the secret message into cipher text to ensure two layer security of the message. This proves that our intended method is feasible. We aim to improve their approach and apply it on web authentication system. Xu et. al [15] proposed an effective LSB based steganographic algorithm which is based on the classic K-means algorithm. A conducted test case shows that the algorithm can hide 60% of the size of the cover without any visual artifacts. Devi and Sharma [16] proposed an improved detection of LSB in gray scale and colour images. Their methods apply only for LSB replacement method and not for LSB matching. We focus on utilizing LSB insertion method based on matching RGB pixels as a better substitution for embedding since this method is robust and also yields and retain high capacities for embedding.

IV. Stega-AES Based Web Authentication Security Method

We are now going to give a detailed implementation of our Steganography AES based Web Authentication Security. As visible in figure 5 our Steganography-AES based Web authentication security begins with the first step which is the acquisition of users Photo ID and since modern mobile devices are transcending from biometric to face recognition for security the acquisition of Photo ID will not be a hassle. Then secondly what is followed is the acquisition of users credential for login into the system of the web application.

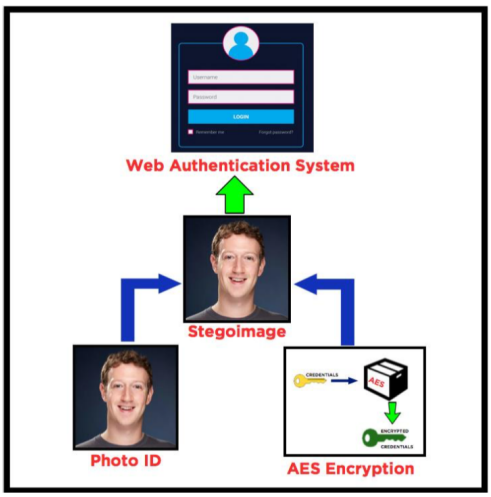


Figure 5. Stega-AES Based Web Authentication System

The last step is to combine the Photo ID with the resulting encrypted AES Encrypted credential to form the Stegoimage which is lastly used to authenticate the user into the web application. The steps can be summarized as follows:

1) **Step 1** - Obtain users Photo ID usually from a webcam or cellphone front facing camera.
2) **Step 2** - Obtain login credentials and perform 128bit AES Encryption on them
3) **Step 3** – Combine both the obtained Photo ID and AES Encryption to form the stegoimage.
4) **Step 4** – Lastly Inspect the stegoimage and if everything is fine then authenticate user.

To elaborate further on our 128 bit AES algorithm which is used in step 2 the obtained login credentials of the user are further encrypted with a robust 128 bit AES encryption as seen in figure 6 below. The produced encrypted credentials are further sent to be combined with the stegoimage in step 3 to produce the desired stegoimage.

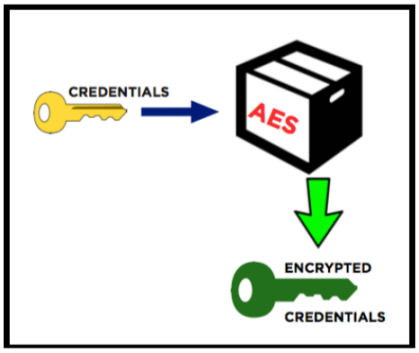


Figure 6. 128 bit AES Encryption Step

Assuming the users login password was $\backslash @Password1972$" then our 128 bit encryption algorithm produces the encrypted password as $\backslash KxxMm + IGTvgL33h7fCPVtQ == $"

which is more robust and hard to decipher. Finally this ciphered password is then translated into binary representation visible in figure 6.1 below to be embedded into the Least Significant Bits of the stegoimage.

01001011	01111000	01111000	01001101	01101101	00101011
01001001	01000111	01010100	01110110	01100111	01001100
00110011	00110011	01101000	00110111	01100110	01000011
01010000	01010110	01110100	01010001	00111101	00111101

Figure 6.1 Binary Representation of Password

The method for encrypting the above message can be given as follows in Java:

```
1. public static String AESencrypt(String
    strToEncrypt){
2. try{
3. Cipher cipher =
    Cipher.getInstance("AES/ECB/PKCS5Padding");
4. cipher.init(Cipher.ENCRYPT_MODE, secretKey);
5. setEncryptedString(Base64.encodeBase64String
(cipher.doFinal(strToEncrypt.getBytes("UTF-8"))));
6. }
7. catch (Exception e)
8. {
9. System.out.println("Error! could not complete
    encryption: "+e.toString());
10. }
11. return null;
```

The algorithm simply instantiates a cipher object which is used to get an AES instance of the *secretkey* and this is invoked in line 4 of the code listing. Then a Base64 encoding is applied for the returned bit stream which is encoded using the UTF-8 ISO standard. Lastly, an exception is caught in cases where the *AESencrypt* method fails. The *AESencrypt* method can also be translated into PHP or JavaScript for web applications. We have tested this and the method works as expected for both domains.

V. Experimental Results and Discussion

A. Steganalysis Results

To test the robustness of our security we have chosen to test the final stegoimage obtained after step 3 for robustness. We chose to do this to test if the steganography is not prone to steganalysis attacks.

1) Least Significant Bit Enhancement

Least significant bit enhancement algorithm extracts the LSB of each pixel (or LSB of red, green and blue value for RGB images). The algorithm sorts out the pixels and If LSB = 1 then set pixel to 255 else If LSB = 0 then set pixel to 0. Each colour can be interpreted as a point in a three-dimensional space, and this forms an RGB colour cube as displayed in figure 7 below. The steganography algorithm automatically creates copy of the sorted pallete of which it can then manipulate and hide messages out. Sorting is usually done according to priority variables such as luminance and etc.

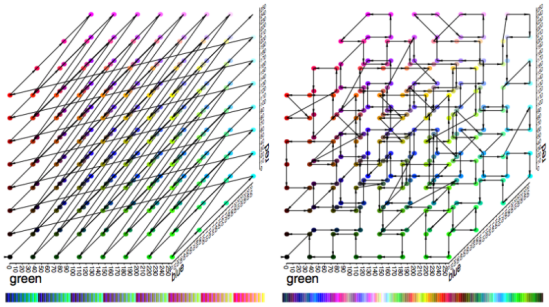


Figure 7. RGB Pixel order Pallete

For our test case we have obtained two images that both look the same anesthetically after embedding the encrypted messages and the images are as Follows:

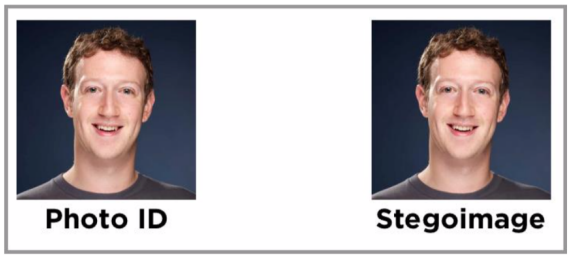


Figure 8. Steganography Image files

As seen in figure 8 both images look unsuspicious and healthy. Also the overall change in size has not drastically increased to a level which compromises its security. Before embedding our cover size was 29KB and after embedding it was 300KB which is still the normal range for such images. The obtained LSB Enhancement from the above images displayed in figure 9 is as follows.

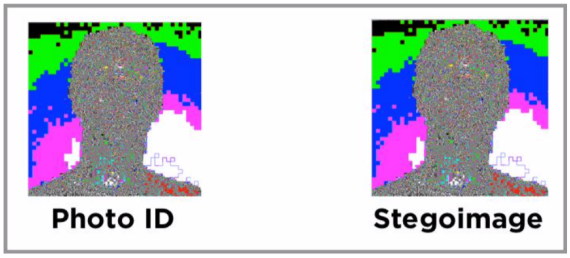


Figure 9. LSB Enhnacement Applied

For the normal short messages and passwords this does not look suspicious and is absolutely fine. However, if we embed messages of large sizes then this will start to show distortion on the upper part of the image after embedding a very large message.

2) Average LSB and Chi Squared

The average LSB and Chi Squared distribution indicates the peak frequencies distribution of a Steganogram. When

embedding on the least significant bits of an image the least significant bits transform values and only differ in the least significant bit. What results are what is known as pairs of value. If the least significant bits are equally distributed, the frequencies of both values of the pairs of value become equal provided the bits used for the LSB are equally distributed.

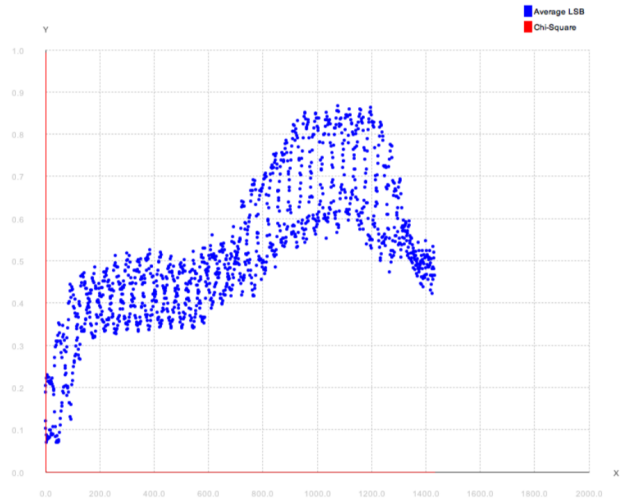


Figure 10. Average LSB and Chi squared distribution

From figure 10 we see a normal distribution on our stegoimage which is totally identical to Average LSB distribution of the Photo ID image before embedding. The Chi squared is not visible because of the low ratio of the experienced distortion on the Least Significant Bits. If this ratio is high then the Chi-Square peak will be high as visible in figure 11 below. If this ratio is low and not high the steganography will be robust and none suspicious and information will be secured.

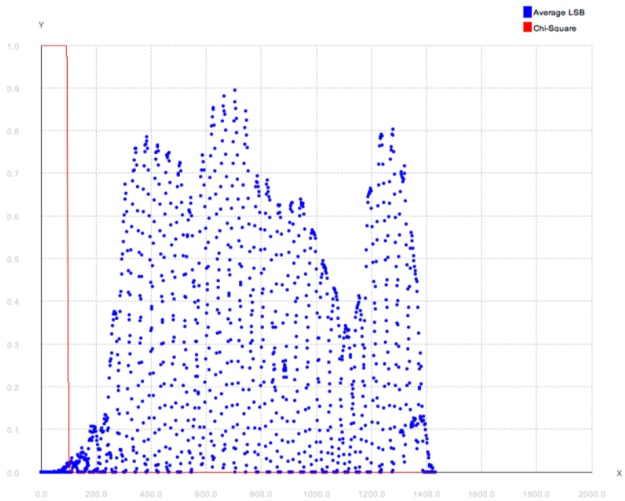


Figure 11. Average LSB and Chi Square Distortion

We have also tested our stegoimage by embedding an image payload into the final stegoimage and the Average LSB distribution and Chi Squared reflected anomalies and constant high peaks visible in figure 11 which can be easily suspected.

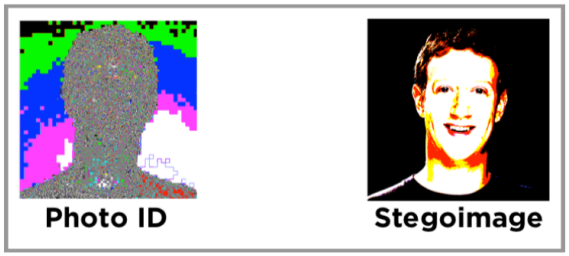


Figure 12. LSB Enhancement image payload

What is visible from figure 12 is a clear vulnerability which tell us that image payload increase the overall frequency of the image by a large factor hence why after the LSB Enhancement the payload was compromised and the stegoimage displayed the image it was embedding although in low ration however, this will be valuable for hackers and etc.

VI. Conclusion and Future Work

In this paper we have introduced and implemented a Web Authentication Security Using Image Steganography and AES Encryption. We have fully demonstrated how it implements 128 bit AES Encryption for robust security. We have also showed that our method is able to handle embedding user credentials which are first encrypted using AES before embedded into the stegoimage that will be used as authentication. Lastly we performed steganalysis attacks which included the LSB Enhancement method, Average LSB and Chi- squared to our produced stegoimages. The results show that if payloads are small in size then LSB enhancement cannot detect the payloads in the stegoimages. However, when the payload is large then high distortion will be experienced and an LSB Enhancement attack alone will be able to detect that it has been concealed using steganography. Furthermore our results indicate that using images as payloads in the final stegoimages does not perform well for our Web Authentication Security system. In future we plan to research and alleviate this problem so that images can be used as payloads without compromising the system's security at all.

References

[1] S. Suri, H. Joshi, V. Mincoha, and A. Tyagi, "Comparative analysis of steganography for coloured images," *International Journal of Computer Sciences and Engineering*, vol. 2, no. 4, pp. 180–184, 2014.

[2] S. Dey, K. Mondal, J. Nath, and A. Nath, "Advanced steganography algorithm using randomized intermediate qr host embedded with any encrypted secret message: Asa_qr algorithm," *International Journal of Modern Education and Computer Science*, vol. 4, no. 6, p. 59, 2012.

[3] A. Cheddad, J. Condell, K. Curran, and P. Mc Kevitt, "Digital image steganography: Survey and analysis of current methods," *Signal processing*, vol. 90, no. 3, pp. 727–752, 2010.

[4] J. Nath, S. Das, S. Agarwal, and A. Nath, "Advanced steganographic approach for hiding encrypted secret message in lsb, lsb+ 1, lsb+ 2 and lsb+ 3 bits in non standard cover files," 2011.

[5] V. L. Reddy, A. Subramanyam, and P. C. Reddy, "Implementation of lsb steganography and its evaluation for various file formats," *Int. J. Advanced Networking and Applications*, vol. 2, no. 05, pp. 868–872, 2011.

[6] G. Nehru and P. Dhar, "A detailed look of audio steganography techniques using lsb and genetic algorithm approach," *International Journal of Computer Science Issues (IJCSI)*, vol. 9, no. 1, p. 402, 2012.

[7] A. Nag, S. Biswas, D. Sarkar, and P. P. Sarkar, "A novel technique for image steganography based on block-dct and huffman encoding," *arXiv preprint arXiv:1006.1186*, 2010.
[8] A. M. Al-Shatnawi, "A new method in image steganography with improved image quality," *Applied Mathematical Sciences*, vol. 6, no. 79, pp. 3907–3915, 2012.
[9] Y. Lim, C. Xu, and D. D. Feng, "Web based image authentication using invisible fragile watermark," in *Proceedings of the Pan-Sydney area workshop on Visual information processing-Volume 11*. Australian Computer Society, Inc., 2001, pp. 31–34.
[10] N. Jain, S. Meshram, and S. Dubey, "Image steganography using lsb and edge–detection technique," *International Journal of Soft Computing and Engineering (IJSCE) ISSN*, vol. 223, 2012.
[11] J. Lu, F. Liu, X. Luo, and X. Song, "Steganalysis of lsb matching based on image noise," in *Multimedia and Ubiquitous Engineering*. Springer, 2014, pp. 211–217.
[12] X. Li, T. Zhang, Y. Zhang, W. Li, and K. Li, "A novel blind detector for additive noise steganography in jpeg decompressed images," *Multimedia tools and applications*, vol. 68, no. 3, pp. 1051–1068, 2014.
[13] E. Walia, P. Jain, and N. Navdeep, "An analysis of lsb & dct based steganography," *Global Journal of Computer Science and Technology*, 2010.
[14] M. R. Islam and M. AyashaSiddiqa, "Palash uddin, ashis kumar mandal and md. delowar hossain,"an efficient filtering based approach improving lsb image steganography using status bit along with aes cryptography" in proceedings of 3rd international conference on informatics," *Electronics & Vision*, pp. 1–6, 2014.
[15] S. Xu and S. Lai, "An optimal least significant bit based image steganography algorithm," in *Proceedings of International Conference on Internet Multimedia Computing and Service*. ACM, 2014, p. 42.
[16] M. Devi and N. Sharma, "Improved detection of least significant bit steganography algorithms in color and gray scale images," in *Engineering and Computational Sciences (RAECS), 2014 Recent Advances in*. IEEE, 2014, pp. 1–5.

Performance Analysis of Machine Learning Classifiers for Intrusion Detection

Skhumbuzo Zwane
Department of Computer Science
University of Zululand
KwaDlangezwa, South Africa
goodwillzwane@gmail.com

Paul Tarwireyi
Department of Computer Science
University of Zululand
KwaDlangezwa, South Africa
tawireyiP@unizulu.ac.za

Matthew Adigun
Department of Computer Science
University of Zululand
KwaDlangezwa, South Africa
adigunM@unizulu.ac.za

***Abstract*—Modern tactical wireless network (TWN) communication technologies are not only capable of transmitting voice but also capable of transmitting data. Due to such capabilities, TWN have high security requirements as any security breach can lead to detrimental effects. Hence, securing such an environment is not only a requirement but also a virtual prerequisite to the network centric warfare operational (NCW) theory. One key to securing this environment is to promptly and accurately recognize information warfare attacks directed to the network and respond to them. This is achieved using intrusion detection systems (IDS). However, false detection of nodes in hostile environment remains a major problem that need to be addressed. Recently, machine learning methods and algorithms have shown applicability and are growing research area for cyber security and intrusion detection. Conversely, several decades of research in the field of machine learning have resulted in a multitude of different algorithms for solving a broad range of problems. The question then becomes, which one amongst these machine learning algorithms have the potential to enhance or address IDS issues in TWN. In this paper, seven machine learning classifiers are analyzed; Multi-Layer Perceptron, Bayesian Network, Support Vector Machine (SMO), Adaboost, Random Forest, Bootstrap Aggregation, and Decision Tree (J48). WEKA tool was used to implement and evaluate the classifiers. The results obtained indicate that ensemble-based learning methods outperformed single learning methods when we consider the detection accuracy metrics; AUC, TPR, and FPR. However, ensemble classifiers tend to be slower in in terms of build time and model test time.**

Keywords—Tactical Network, ensemble learning, Machine learning, build-time, test-time

I. INTRODUCTION

Tactical Wireless Networks (TWN) have revolutionise military communications and operations. TWN have changed the way military forces communicate, disseminate intelligence, issue orders, and report about their missions. At the current stage, military communication systems are not only capable of transmitting voice but also capable of transmitting data [1]. TWNs are characterised by limited power, low memory budget, ad hoc, dynamic, usually contested by adversaries, and operate in hostile environments [2].

In TWN, security is not only a requirement but a virtual prerequisite to the network centric warfare (NCW) operational theory [2]. Researchers on TWN security have implemented different mechanisms to address security requirements, for example, smart jammer mitigation [3] and Extremely Lightweight Intrusion Detection (ELIDe) [4]. Intrusion detection systems have been recently adopted in TWNs to detect network intruders. For example, detect intruders by their misbehaviors, such as dropping or modifying packets [2]. However, the environment in which TWNs are deployed present challenges to the intrusion detection methods. For example, in a hostile environment, packets are regularly dropped and corrupted, detection mechanisms see this behavior as malicious and false detection occur. Further, false detection of hostile nodes in hostile environment is a major issue that need to be addressed [2].

Recently, Machine learning (ML) based intrusion detection methods have shown applicability and promising results in addressing cyber security issues. The present paper conduct a performance evaluation of seven ML algorithms. The aim is to analyse and determine which ML algorithm can help address false detection problem in TWN. The algorithms are evaluated in terms of; true positive rate, false positive rate, area under ROC Curve, build time, and test time. To address false detection problem an IDS should be able to detect intruders in real-time with high accuracy, and also retraining the model should take the minimum time to ensure the effectiveness of the intrusion detection technique in TWN. The ML algorithms are implemented and evaluated using Waikato Environment for Knowledge Analysis (WEKA) and TPR, FPR, AUC, build time, and test time are measured.

The rest of the paper is organized as follows. Section II provide the literature review and evaluation datasets. A brief overview of the implemented classifiers is presented in Section III. Section IV contain the experiment and results. Finally, the conclusions are provided in Section V.

II. LITERATURE REVIEW

[5] defined intrusion detection as the process of identifying computing or network activity that is malicious or unauthorised. Intrusion detection systems (IDS) and other mechanisms have been used extensively during the years to achieve security in Tactical Wireless networks (TWN) [2] [3] [4]. The work of [4] proposed Extremely Lightweight Intrusion Detection (ELIDe), an intrusion detection system with the capacity to approximate snort-like signature matching against both inbound and outbound network traffic in a single host. ELIDe uses a linear machine learning classifier which relies on snort implementation as its training oracle. This method achieved snort-like capabilities while using less than 2% of the peak memory footprint demanded by snort [4]. The authors argued that ELIDe was suitable for

operation in constrained environment like TWN. While this is true, their method approximates the functionality of a signature based IDS, Snort. In today's rate of development in technology, anomaly-based intrusion IDS are preferred over signature-based IDS.

The work of [3] proposed an adaptive rapid channel-hoping scheme to mitigate smart jammers. Their method used Dwell window and Deception mechanisms to achieve its functions. Even though the scheme proved to be effective, its capabilities are limited to only defending against smart jammers.

In a survey by [2], the authors reported that some security, availability, and reliability issues have been solved in TWN. However, false detection of hostile nodes by implemented IDS in hostile environment is still an open research gab that needs to be addressed.

A. Machine Learning IDS

ML techniques have been around for a long time, however, finding a way to use them efficiently and in real-time is a new trend [6]. This is due to the emergence of new computing technologies and the availability of data. Recent approaches in intrusion detection have applied machine learning to improve detection rate [7] [8] [9]. Conversely, due to the large number of available ML algorithm, it remains unclear which algorithms are suitable for specific intrusion detection tasks. [10] Compared the effectiveness of five machine learning models; Random Forest, J48, Support Vector Machine, CART, and Naïve Bayes. They observed that Random Forest outperformed the other algorithms by achieving higher accuracy than the others. Even though random forest performed better, the authors did not report how long it takes to build and test each of their models. [11] compared Naïve Bayes, Bayesian Network, Random Forest, Multi-Layer Perceptron (MLP), and SOM for intrusion detection in computer networks. The KDD-CUP 99 datasets were used to build the detection models. The authors reported that MLP achieved higher detection accuracy. They also reported that the process of building the model took about 12 hours, which is regarded as unacceptable for dynamic and fast changing networks.

In the mentioned studies, we observed that the authors selected their best performing classifier by considering the detection and accuracy metrics. However, the training time and the time to classify a new instant were not considered. The work of [12] argued that the training time is an important factor due to the changing cyber-attack types and features. In addition, anomaly detectors used in TWN need to be frequently trained with fresh malware signatures. In addition, the time for classifying a new instant reflect the reaction time and the packet processing power of an IDS. Hence, their findings are not applicable to TWN IDS implementation.

B. Evaluation Datasets

Many studies in literature employ the KDD-CUP 99 and NSL-KDD for evaluating IDSs. The works of [13] and [14] argued that the KDD-CP 99 datasets contain a large number of redundant records in the training set. There are also reports of multiple missing records which are a factor in changing the nature of the data. In addition, in the NSL-KDD datasets which is an improved version of the KDD-CP 99 dataset, [15] claimed that the dataset does not comprehensively represent modern low footprint environment. To address those limitations, other datasets are being proposed as benchmark dataset for IDS evaluation [13] [14]. These datasets ameliorate the shortcomings of KDD-CP 99 datasets, such as they represent modern network behaviours.

In this paper, the UNSW-NB15 network datasets [15] are utilized. The UNSW-BN15 network dataset is a hybrid of modern normal and abnormal network traffic created using the IXIA PerfectStorm tool at the Cyber range lab of the Australian Centre for Cyber Security (ACCS) [15]. The dataset is made up of nine different families of attack instances, the different attack categories are shown in Table I. A brief description of how the dataset was generated and list of all the generated features can be found in the article by Moustafa, et al [15].

TABLE I. UNSW-NB15 NETWORK DATASET DESCRIPTION

Type	Number of Instances	
	Training	***Testing***
Normal	37 000	56 000
Fuzzers	6 062	18 184
Backdoors	583	1 746
Analysis	677	2 000
DoS	4 089	12 264
Exploits	11 132	33 393
Generic	18 871	40 000
Reconnaissance	3 496	10 491
Shellcode	378	1 133
Worms	44	130
Total	82 332	175 341

III. OVERVIEW OF MACHINE LEARNING CLASSIFIERS

This section provides a general overview of the evaluated machine learning methods for intrusion detection in TWN.

The task of intrusion detection can be seen as a classification problem where the IDS classify between malicious or normal network behaviour [18]. In this work, the interest is in the following machine learning classifiers;

A. Multilayer Percepton Classifier

Multi-layer Perceptron (MLP) Classifier is one of the most common functions classifiers. MLP has proven to be effective in several application areas, which include time series, classification, and for regression problems [16]. MLP algorithm can be implemented with various transfer functions, e.g. Hyperbolic, Linear, and Sigmoid functions. The most important design considerations of a MLP algorithm is the number of expected outputs and the number of hidden layers [17]. Studies show that MLP usually achieve above 90% detection accuracy. In this paper, the default configurations for MLP in WEKA were used.

B. Support Vector Machine Classifier

Support Vector Machine (SVM) is a hyperplane that separates a set of positive examples from a set of negative

examples with maximum margin. Sequential Minimal Optimization (SMO) algorithm is used for training a support vector classifier using polynomial or RBF kernels [18]. This method is useful when classifying non-linear datasets. For the implementation of SVM, we have used WEKA's SMO since our datasets contained non-linear relations.

C. J48 Classifier

J48 Classifier belongs to the decision tree algorithm family. The J48 classifier was designed to improve the C.4.5 algorithm [19]. The algorithm's expected output is in the form of decision trees but with more stability between computation time and accuracy [17]. We used WEKA to implement the J48 classifier, in its default settings.

D. Random Forest Classifer

Random Forest Classifier is an ensemble classification algorithm which uses trees as base classifiers [20]. In the training phase, some data may be used more than once in the training of the classifier, while others might never be used. Random forest can be perceived as a group of trees fitted to different subset of the data. This improve detection rate of traditional decision tree classifier. From WEKA, we apply random forest to the described dataset, with its default base classifier (decision tree).

E. Bayesian Network Classifier

Bayesian Network (BN) classifier can be implemented by optimizing of log-likelihood (LL) also known as generative learning. The goal of this paradigm is to obtain parameters characterizing the joint distribution in the form of local conditional distributions and then estimate class-conditional probabilities using Bayes rule. Alternatively the conditional-log-likelihood (CLL) can be optimized. The goal here is to directly estimate the parameters associated with the class-conditional distribution [21]. This classifier is implemented using WEKA, at its default configuration.

F. Adaptive Boosting Classifier

Adaptive Boosting (Adaboost) Classifier is an ensemble classifier, its core function is to fit a sequence of weak learners, such as, models that are slightly better than random guessing, on repeatedly modified versions of the data. The predictions from all of them are then combined through weighted majority vote (or sum) to produce the final prediction. In our implementation, the default WEKA Adaboost settings are adopted.

G. Bootstrap Aggregation Classifier

Bootstrap Aggregation (Bagging) Classifier is an ensemble method that creates individuals for its ensemble by training each classifier on random redistribution of the training set. Each classifier's training set is generated by randomly drawing, with replacement from the training set. Bagging classifier is implemented using the default WEKA configurations.

IV. Experiments and Results

As mentioned in above sections, WEKA tool is adopted for this experiment. WEKA version 3.8 installed in Windows 10, Intel(R) Core i7-6700 CPU @ 3.40GHz machine, with 8 GB RAM, is used for this experiment.

A. Evaluation Metrics

Seven machine learning classifiers are evaluated for suitability in TWN. To assecess the suitability, three accuracy-based metrics and two time based metrics are adopted:

- True Positive Rate (TPR) – this metric corresponds to the proportion of positive data points that are correctly classified as positive, with respect to all positive data points.
- False Positive Rate (FPR) – this metric corresponds to the proportion of negative data points that are mistakenly considered as positive, with respect to all negative data points.
- Area Under ROC Curve (AUC) – combine both FPR and TPR into one metric. TPR and FPR are computed with different thresholds, then plot them in a single graph, with the FPR values on the abscissa and TPR values on the ordinate. The resulting curve is called the ROC curve, and the metric we consider is the Area Under the Curve (AUC).
- Build-time – the time it takes for each classifier to build a model using training data
- Test-time – the time it takes to test a model using the testing data

B. Dataset Preprocessing

As mentioned earlier, this study utilizes the UNW-NB15 network datasets. The dataset is loaded to WEKA tool. The ML classifiers are used to build intrusion detection models using a training set with **82,332** instances, and tested using **175,341** unknown instances. Data pre-processing is used to restructure and remove features that may cause overfitting or no information gain. The training and testing sets contain 45 attributes each, which include two types of labels; original ***Class*** **lable**, 0 for normal, and 1 for abnormal. ***attack_Category*** labels each abnormal instance based on attack category, see Table I. Thus, in this experiment, we use pre-processing tools to remove ***attack_Category*** label and the instance **ID** to avoid data overfitting. Our final evaluation dataset contains 43 attributes, and a class label with 0 for normal and 1 for malicious.

C. Evaluation of Machine learning classifiers

Multilayer Perceptron, Bayesian Network, Support Vector Machine (SMO), Adaboost, Random Forest, Bootstrap Aggregation, and Decision Tree (J48) classifiers are employed on UNSW-NB15 labelled dataset using WEKA. The experiment is repeated 6 times for each classifier, and the average is used for consistency purposes.

TABLE II. CLASSIFIER TPR, FPR, AND AUC

Classifier	Experimentation values		
	TPR	***FPR***	***AUC***
Multilayer-Perceptron	0.933	0.142	0.951
Adaboost	0.903	0.087	0.968
Random Forest	0.902	0.057	0.981
Bagging	0.901	0.057	0.952
J48	0.887	0.069	0.952
SMO	0.882	0.073	0.905
Bayesian Network	0.809	0.123	0.965

Table II shows the TPR, FPR and AUC of each classifier model using two class labels (normal and malicious). As cab be seen in Table II, multilayer perceptron missed fewer positive data points, achieving TPR of 0,933. The second is Adaboost with 0,903, random forest at 0,902 and bootstrap aggregation at 0,901. The reminder of the classifiers achieved TPR of less than 90 %.

When we focus on FPR, Multilayer perceptron have higher FPR followed by the Bayesian Network, with 0.142 and 0.123. We observe that the other classifiers have FPR less than 0,1, as shown in Table II.

The AUC is one of the most important metric used to measure classifier performance. Table II shows that all the models achieved AUC above 0.9. However, Bootstrap aggregation have the highest AUC (0.986), followed by Random forest at 0.981, and Adaboost at 0.968. Figure 1 shows a graphical representation of the results in Table II.

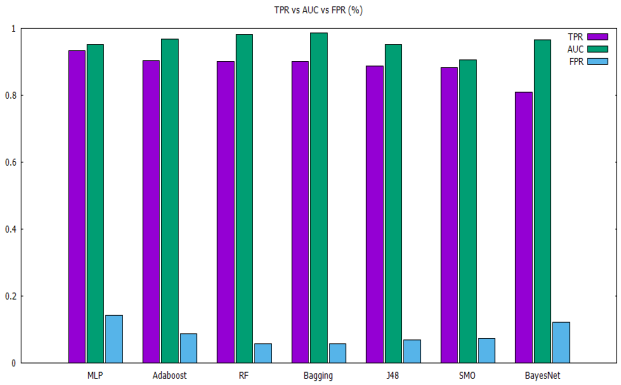


Fig. 1. Classifier TPR, AUC, and FPR

TABLE III. CLASSIFIER BUILD AND TEST TIME

Classifier	Time in seconds	
	Build (sec)	***Test (sec)***
Multilayer-Perceptron	86303.22	14.35
Adaboost	9.02	1.06
Random Forest	28.68	4.91
Bagging	18.86	1.19
J48	8.08	1.02
SMO	541.88	1.44
Bayesian Network	2.56	1.63

The time for training a model is an important factor due to the ever-changing cyber-attack types and features. Table III shows the classifier build and test time. Multilayer-Perceptron (MLP) took longer to build the model, followed by SMO. MLP took 86303.22 seconds, about 24 hours to train the model, this is followed by SMO with an average of 9 minutes (541,88 seconds). Random forest and Bootstrap Aggregation with build time 28,68 sec and 18,86 secs follow. The remainder of the list are all bellow 10 seconds, as shown in Figure 2.

When we consider test time, Multilayer perceptron takes more time, 14,35 seconds, followed by random forest, 4,91 secs. The remainder of the algorithms were able to test their models in below 2 secs, see Table III.

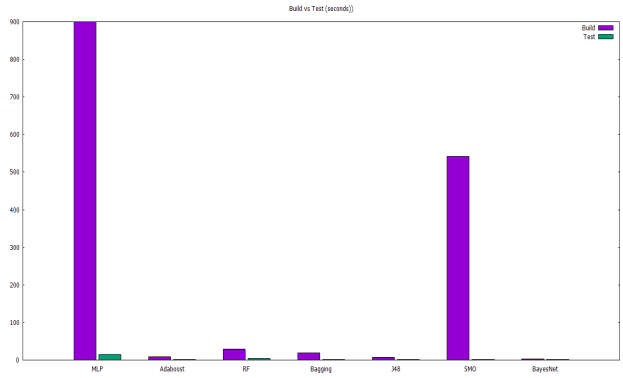


Fig. 2. Classifier build and test time (secs)

D. Analysis and results Orientations

It is important to note that the higher the TPR, means fewer positive data points missed. Therefore, classifiers that achieve higher TPR are preferred over the ones that generate lower TPR in TWN. From the results above, MLP, Adaboost, Random Forest, and Bagging generated higher TPR. These classifiers demonstrate that they have a potential to address the false detection issues in TWN. Hence, a classifier with high TPR is required so to correctly classifier malicious traffic in the network.

The second important performance metric is the FPR. Higher FPR, indicate that more negative data points are misclassified by the classifier. Which means a classifier has a higher chance of producing many false alarms if the FPR is high. Multilayer perceptron have higher FPR followed by the Bayesian Network, this demonstrate that those two classifiers can produce more false positive prediction than the other classifiers. Therefore, using those classifiers for intrusion detection in TWN is not recommended.

Another detection metric used to measure detection performance is the AUC. It should be noted that the higher the AUC the better is the model. As observed in the above subsection, Bootstrap, random forest, and Adaboost achieved higher AUC, hence these classifiers are suitable for implementation for intrusion detection. While TPR and FPR are important, it is important to consider the AUC as the best measure for selecting ML classifier for TWN.

On the other hand, the time to classify a new instance is an important factor that reflects the reaction time and the packet processing power of the intrusion detection system. Also, anomaly detectors need to be trained frequently or incrementally, with fresh malware signature updates for

effective results. From the results we also observe that MLP can present challenges for real-time intrusion detection in TWN. This is due to the large amount of time required to build and test the MLP classifier. From the results, Random Forest, Bagging, Adaboost, and J48 are the best performers in terms of TPR, FPR, and AUC

It should be noted that 3 of these classifiers are ensembled classifiers. In Addition, Adaboost, Bagging, J48, and Bayesian Network classifiers, are much faster in building and testing their model. Here we note that the ensemble classifiers tends to be much slower when building and testing a model.

E. Limitations and Future Works

The machine learning classifiers are evaluated using UNSW-NB15 network datasets. Although these datasets are new and reflect modern network patterns, they do not reflect a hostile environment, with dynamic topology and hash terrain environment.

In future work, ensemble learning classifiers will be evaluated for the task of intrusion detection. This will be achieved using simulated TWN traffic datasets. Furthermore, an investigation on how probabilistic, Rule induction, decision tree, and non-probabilistic based classification algorithms can be ensembled for better performance and efficiency in TWN is required.

V. CONCLUSIONS

The results obtained in this study demonstrate and show that ensemble learning methods perform better than single learning methods when accuracy based metrics are considered. This implies that ensemble machine learning have the potential to address false detection issues in TWN. On the other hand, single learning tends to be much quicker when building and testing a model. Since tactical TWN are deployed in hostile environment, ensemble classifiers can help mitigate the problem of false detection of nodes. However, their limitations are in the amount of time required to build intrusion detection models.

This study have evaluated seven machine learning classifiers using WEKA, respectively. The aim of the study was to conduct a performance analysis of different machine learning classifiers to assess applicability of ML methods in addressing intrusion detection issues in TWN. We recorded and observed TPR, FPR, AUC, build and test time as the major performance metrics for assessing suitability of each classifier for intrusion detection in TWN environment.

ACKNOWLEDGMENT

The authors thank the Centre of Excellence in Mobile e-Services, and Department of Computer Science at the University of Zululand.

REFERENCES

[1] E. Gerhards-padilla and F. Fkie, “Intrusion Detection in Tactical Multi-Hop Networks,” 2009.

[2] W. Pawgasame and K. Wipusitwarakun, “Tactical wireless networks: A survey for issues and challenges,” *2015 Asian Conf. Def. Technol.*, pp. 97–102, 2015.

[3] J. Jeung, S. Jeong, and J. Lim, “Adaptive Rapid Channel-hopping Scheme Mitigating Smart Jammer Attacks in Secure WLAN,” pp. 1231–1236, 2011.

[4] R. J. Chang, R. E. Harang, G. S. Payer, and R. R. E. Harang, “Extremely Lightweight Intrusion Detection (ELIDe),” no. December, 2013.

[5] M. Indhumathi and S. Kavitha, “Distributed Intrusion Detection System for Cognitive Radio Networks Based on Weighted Fair Queuing Algorithm,” vol. 3, no. 4, pp. 426–436, 2018.

[6] M. Zaman, “Evaluation of Machine Learning Techniques for Network Intrusion Detection,” *NOMS 2018 - 2018 IEEE/IFIP Netw. Oper. Manag. Symp.*, pp. 1–5, 2018.

[7] A. L. Buczak and E. Guven, “A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection,” vol. 18, no. 2, pp. 1153–1176, 2016.

[8] R. Vijayanand, D. Devaraj, and B. Kannapiran, “Intrusion detection system for wireless mesh network using multiple support vector machine classifiers with genetic-algorithm-based feature,” *Comput. Secur.*, vol. 77, pp. 304–314, 2018.

[9] T. Dunning and E. Friedman, *Practical Machine Learning: A New Look At Anomaly Detection*. 2014.

[10] S. Revathi and A. Malathi, “A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection,” vol. 2, no. 12, pp. 1848–1853, 2013.

[11] F. Ertam, F. Õ. K. Õ, and O. Yaman, “Intrusion Detection in Computer Networks via Machine Learning Algorithms,” 2017.

[12] C. Kolias, G. Kambourakis, A. Stavrou, and S. Gritzalis, “Intrusion Detection in 802 . 11 Networks : Empirical Evaluation of Threats and a Public Dataset,” no. c, pp. 1–24, 2015.

[13] P. Gogoi, M. H. Bhuyan, and D. K. Bhattacharyya, “Packet and Flow Based Network Intrusion Dataset,” 2012.

[14] E. Harshini, S. Selvakumar, and T. N. State, “SSENet-2011 : a Network Intrusion Detection System Dataset and its Comparison with KDD CUP 99 Dataset,” 2011.

[15] N. Moustafa, J. Slay, and I. Technology, “Intrusion Detection systems,” 2015.

[16] S. K. Pal, S. Member, S. Mitra, and S. Member, “Multilayer Perceptron, Fuzzy Sets, and Classification,” vol. 3, no. 5, pp. 683–697, 1992.

[17] I. M. Alsmadi and A. AlEroud, “SDN-based real-time IDS/IPS alerting system,” in *Studies in Computational Intelligence*, vol. 691, 2017, pp. 297–306.

[18] B. Sch and A. Smola, “Support Vector Machines and Kernel Algorithms,” pp. 1–22, 2002.

[19] J. Ross, Q. Morgan, and K. Publishers, “Book Review : C4 . 5 : Programs for Machine Learning,” vol. 240, pp. 235–240, 1994.

[20] M. A. Jabbar, R. Aluvalu, S. Satyanarayana, and S. Reddy, “ScienceDirect ScienceDirect RFAODE : A Novel Ensemble Intrusion Detection System,” *Procedia Comput. Sci.*, vol. 115, pp. 226–234, 2017.

[21] C. S. Division, M. Park, P. Langley, and P. Smyth, “Bayesian Network Classifiers *,” vol. 163, pp. 131–163, 1997.

Rule-based Control Strategy for a River-based Grid-connected Hydrokinetic System

Sandile Phillip Koko
Department of Electrical, Electronic and Computer Engineering
Central University of Technology
Bloemfontein, South Africa
skoko@cut.ac.za

Kanzumba Kusakana
Department of Electrical, Electronic and Computer Engineering
Central University of Technology
Bloemfontein, South Africa
kkusakana@cut.ac.za

Herman Jacobus Vermaak
Department of Electrical, Electronic and Computer Engineering
Central University of Technology
Bloemfontein, South Africa
hvermaak@cut.ac.za

***Abstract*—This paper presents an optimal control algorithm for improving the reliability of a river-based grid-interactive hydrokinetic system. The aim is to guarantee a reliable system operation during load demand uncertainties. The proposed system consists of a micro-hydrokinetic pumped-hydro storage (MHK-PHS) system, modelled to supply electricity to a commercial load profile. The open-loop optimization-based approach resulted in an unmet load demand and excessive energy supply, due to load demand uncertainty. Hence, a rule-based control algorithm is presented to ensure that the load demand is reliably met at all times. Thus, the results prove that the developed rule-based control algorithm, allowed the commercial load demand to be reliably met by solving both the unmet load demand and excessive energy supply challenges.**

Keywords—Micro-hydrokinetic, pumped-hydro storage, time-of-use, commercial load, rule-based control.

I. INTRODUCTION

The global issues of climate change, the ever-increasing energy demand and depletion of fossil fuel reserves necessitate the use of renewable energy (RE) sources [1-2]. The construction of additional fossil fuel power plants (FFPPs) is not an economical solution due to the ever-increasing price of fossil fuels [3]. During peak demanding hours, electric grid operators encounter temporal challenge of an imbalance between the demand and the supply [4]. This leads to the introduction of the variable electricity pricing strategies by the grid operators, in order to control/reduce the peak demand. Time-of-use (TOU) pricing mechanism is the most widely adopted mechanism by grid operators around the world [4-6].

In contrary to the construction of additional FFPPs, the use of RE sources is the key to mitigate all of the above-mentioned challenges. Among various RE technologies, hydrokinetic technology has gained a considerable attention since it is easily predictable. It generates electricity by making use of the river streams, tidal current or artificial water channels with flowing water speed of 0.5 m/s or above [7-8]. It has proved to off-set the stochastic nature of solar and wind turbine technologies. Various studies have proved that it generates electricity cheaper and better than solar and wind turbine technologies [9-10]. The inclusion of a pumped-hydro storage (PHS) make it more feasible [11].

Allowing consumers to have their own onsite grid-interactive RE system may minimize the chances of demand-supply imbalance during peak demanding periods. However, the challenge is to develop an optimal energy management system instead of expecting consumers to monitor and response to the price fluctuation in real time. The number of recent studies based on optimal energy management have been carried. The conventional open-loop optimization approaches have been used to determine an optimal scheduling. Numerous optimal control studies ignore the load demand uncertainties by assuming that a load demand is constant for all days of the year. Load demand uncertainty might negatively affect the system power flow operation if not well managed. Therefore, this study aims to develop a rule-based control strategy that is capable of handling load demand uncertainties, for a river-based grid-interactive micro-hydrokinetic pumped-hydro storage (MHK-PHS) system, under TOU tariff scheme. The configuration of the proposed system is described in Section II, modeling of RE sources in Section III, objective function and constraints involved in Section IV, the operation and effectiveness of the developed rule-based control algorithm in Section V and VII, respectively. Finally, Section VIII concludes the paper.

II. SYSTEM CONFIGURATION

The configuration of the grid-interactive RE system considered in this work, is as illustrated in Fig. 1. The system comprises of the river-based micro-hydrokinetic system (MHK) for RE generation as well as the pumped-hydro storage (PHS) system for energy storage. The commercial load demands electricity from the three sources, namely the MHK system $P_{1(t)}$, PHS system $P_{2(t)}$ and/or utility grid $P_{3(t)}$. The PHS system can store energy from the MHK system $P_{4(t)}$ and/or from the utility grid $P_{5(t)}$. It can also be seen that the power from the MHK system $P_{6(t)}$ and from the PHS system $P_{7(t)}$ can be sold into the utility grid. The PHS system consist of two separate water channels, for refilling the upper reservoir and for generating electricity, respectively. Hence, both charging and discharging can simultaneously take place.

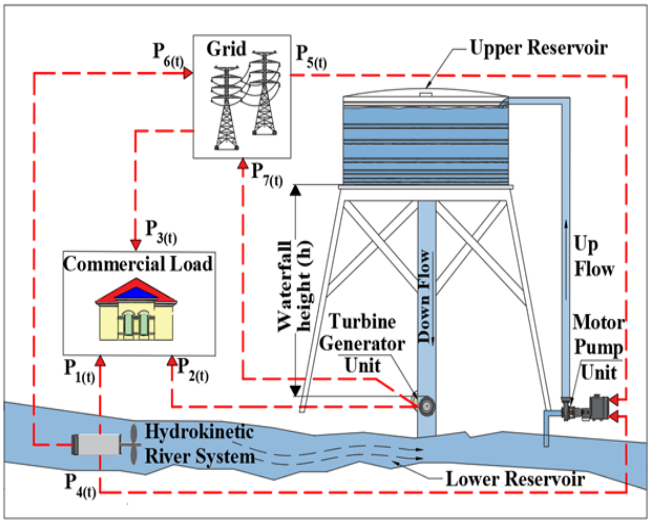


Fig. 1. Schematic diagram of the river-based MHK-PHS system

III. MODELING OF THE ENERGY SYSTEMS

Different components of the proposed river-based MHK-PHS system are mathematically presented as discussed below.

A. Hydrokineitc Model

Hydrokinetic turbines convert the kinetic energy of the flowing river water into rotational energy. The energy produced by the hydrokinetic turbine is calculated as follows [11-12]:

$$E_{HK} = 0.5 \times \rho_W \times A \times v^3 \times C_p \times \eta_{HKT-G} \times t \quad (1)$$

ρ_W is the water density (1000kg/m^3), A is the swept area of the turbine (m^2), v is the water velocity (m/s), C_p is the coefficient of the modelled turbine performance, η_{HKT-G} is the overall efficiency of the hydrokinetic turbine-generator unit and t is the time (s).

B. Pumped-Hydro Storage Model

The quantity of the stored energy in the upper reservoir is proportional to the stored water volume as well as the waterfall height. Hence, the stored potential energy (kWh) is expressed as follows [13-14]:

$$E_{pot} = \frac{V \times \rho_W \times g \times H \times \eta_{T:G}}{3.6 \times 10^6} \quad (2)$$

V is the volumetric storage capacity of the upper reservoir (m^3), g is the gravitational acceleration ($9.81\ m/s^2$), H is the water-head height (m) and $\eta_{T:G}$ is the overall efficiency of the turbine-generator unit of the PHS system.

IV. PROBLEM FORMULATION

The specific objective function as well as the system constraints are implemented from the schematic layout shown in Fig. 1. In this study, the main optimization objective is to minimize electricity bills and to maximize the energy sales revenue, while ensuring zero chances of the unmet load demand. The proposed optimization model is discussed below. Due to the availability of both linear and non-linear constraints shown below, the optimization problem is solved through the use of *fmincon* solver available in *MATLAB* toolbox.

A. Objective Function

The reduction of electricity bills is achieved by lowering the grid power consumption (for both pumping and supplying of the primary load). Maximized energy sales revenue is achieved by selling excess or stored RE to the grid, based on the most profitable TOU periods. Hence, the discrete objective function (F) at any sampling interval (j) is then expressed as follows:

$$F = \sum_{j=1}^{N} C_j \cdot (P_{3(j)} + P_{5(j)}) \cdot \Delta t - C_j \cdot R_a \sum_{j=1}^{N} (P_{6(j)} + P_{7(j)}) \cdot \Delta t \quad (3)$$

N is the total number of sampling intervals, Δt is the sampling time, Cj is the TOU electricity prices at the j^{th} sampling time per kWh and R_a is the feed-in-tariff (FIT) agreement ratio between the consumer and the electric grid operator (assumed to be 65% of the utility retail price in this study).

B. Equality Power Constraints

The commercial load demand must be met by the MHK system, PHS system and/or the grid depending on the TOU tariffs at that point in time. This is discretized as shown by (4). The upper reservoir's state of charge at the beginning of the simulation horizon is allowed to be same at the end of the simulation horizon, as shown by (5). The aim is to allow the repeated implementation of the optimal energy control [15].

$$P_{Load(j)} = P_{1(j)} + P_{2(j)} + P_{3(j)} \quad (1 \le j \le N) \quad (4)$$

$$\sum_{j=1}^{N} (P_{4(j)} + P_{5(j)}) - \sum_{j=1}^{N} (P_{2(j)} + P_{7(j)}) = 0 \quad (1 \le j \le N) \quad (5)$$

C. Inequality Power Constraints

The overall power generated by the MHK system is mainly used to supply the commercial load, the motor pump unit and/or to be sold into the grid, as shown by (6). The power from the MHK system and/or the grid can be stored into the upper reservoir by supplying up to the rated input power of the motor-pump unit ($P^{rated}_{M:P(j)}$), as revealed by (7). The turbine-generator unit can supply up to a maximum of its rated output power ($P^{rated}_{T:G(j)}$) to the commercial load and/or sell into the grid, as shown by (8).

$$P_{1(j)} + P_{4(j)} + P_{6(j)} \le P^{\max}_{MHK(j)} \quad (1 \le j \le N) \quad (6)$$

$$P_{4(j)} + P_{5(j)} \le P^{rated}_{M:P(j)} \quad (1 \le j \le N) \quad (7)$$

$$P_{2(j)} + P_{7(j)} \le P^{rated}_{T:G(j)} \quad (1 \le j \le N) \quad (8)$$

D. Storage Constraints

The dynamics of the storage system is presented by (9). The upper reservoir is refilled using the MHK and/or the grid power. The discharging process takes place when the PHS system supplies power to the load and/or sells energy to the grid. To avoid both over-discharging and floating of the water, the storage level is limited using (10).

$$Cap_{(j)} = Cap_{(0)} + \sum_{j=1}^{N} [(P_{4(j)} + P_{5(j)}) \cdot \frac{\eta_{M:P}}{E_{pot}} \cdot \Delta t] - \sum_{j=1}^{N} (P_{2(j)} + P_{7(j)}) \cdot \frac{\Delta t}{\eta_{T:G} \cdot E_{pot}} \quad (1 \le j \le N) \quad (9)$$

$$Cap^{\min} \le Cap_{(j)} \le Cap^{\max} \quad (1 \le j \le N) \quad (10)$$

Cap^{min} and Cap^{max} are the minimum and maximum allowable capacities, $Cap_{(0)}$ is the initial water level state and $\eta_{M:P}$ is the overall efficiency of the motor-pump unit.

E. Non-linear Constraint

Grid power consumption is only allowed if the load demand cannot be met by both the RE system and the storage system. It can also be consumed if it is economical during off-peak periods. Hence, the model should not tolerate the simultaneous selling and buying from the utility grid as revealed by (11).

$$(P_{3(j)} + P_{5(j)}) * (P_{6(j)} + P_{7(j)}) = 0 \quad (1 \le j \le N) \quad (11)$$

F. Boundary Conditions

All of the above mentioned power control variables are limited between the lower and upper boundaries. They are limited to a minimum boundary limit of zero to the upper boundaries of their maximum or the rated capacity of the system they are supplying, as shown by (12) and (13).

$$0 \le P_{i(j)} \le P_{i(j)}^{\max} \quad (i=1, 4, 6) \quad (1 \le j \le N) \tag{12}$$

$$0 \le P_{i(j)} \le P_{i}^{rated} \quad (i=2, 3, 5, 7) \quad (1 \le j \le N) \tag{13}$$

V. Rule-Based Control Algorithm

The formulated open-loop optimization-based algorithm discussed in section IV, relies heavily on the prediction precision. It works best when the predicted load demand is the same as the actual load demand. In real-time, there is a high probability of load forecasting errors. Therefore, it is practically impossible for the predicted load demand to be equal to the actual load demand throughout the day.

In this study, a rule-based control algorithm is developed to assist the developed open-loop optimization approach in handling load demand uncertainties. The reason for selecting a rule-based control algorithm, is that it does not require forecasting, since the real-time measurements may be used [16]. It behaves like a real-time supervisory control. Hence, it is used to guarantee that the load demand is reliably met at all times, in spite of the load demand uncertainties. This should be achieved without compromising the upper and lower boundaries of the system variables. The flowchart of the developed rule-based algorithm is as shown in Fig. 4. The predicted power variables are presented using the subscript *P*, while the actual ones are without the subscript *P*.

From Fig. 1, it has been observed that the load demand is met by the three power sources. When the predicted load demand (P_{PLoad}) is the same as the actual load demand (P_{Load}), all the power flow variables determined by the open-loop optimization approach are allowed to remain unchanged. Whenever P_{Load} is less than P_{PLoad}, the control algorithm should be able to reduce the excessive power supplied to the load. The first priority is to reduce the predicted grid-to-load (P_{P3}) power while maintaining the power balance of the three sources.

Whenever P_{Load} is greater than P_{PLoad}, the developed rule-based algorithm should economically supplement the unmet demand by increasing the predicted MHK-to-load (P_{P1}) power as a first priority. Therefore, the predicted grid-to-load (P_{P3}) power should be increased as a last priority, only if P_{Load} is still unmet. The overall power supplied to recharge the PHS system is controlled and maintained constant to ensure that the predicted upper reservoir state of charge is the same even after applying the rule-based control algorithm.

VI. Simulation Data

As mentioned before, the optimization problem presented in section IV is solved through the use of the *fmincon* solver in MATLAB. A 9 day commercial load data to be used, has been extracted from an annual load data created by HOMER using a typical commercial load profile found in [17]. A load data from Sunday (1st January 2017) to Monday (9th January 2017) is used to model the predicted load demand, while the demand data from Sunday (16th January) to Monday (24th of January) is used to represent the actual load demand as shown in Fig. 5.

The flowing water speed from a typical river of South Africa has been used to model the hydrokinetic resource as shown in Fig. 2 [17]. Two 1.5 kW hydrokinetic turbines have been used to generated 3 kW output power, when the water speed is above 2 m/s [18]. Low demand season TOU tariff rates and schedule have been used during simulations as shown by (14) and Fig. 3, respectively [19]. The overall simulation parameters are as shown in Table I.

$$C(t) = \begin{Bmatrix} C_p = ZAR1.07/kWh \\ C_s = ZAR0.74/kWh \\ C_o = ZAR0.47/kWh \end{Bmatrix} \tag{14}$$

C_P is the tariff price during peak periods, C_S is the tariff price during standard periods, C_O is the tariff price during off-peak periods and *ZAR* being the South African currency.

TABLE I. Overall Simulation Parameters

Item	Value
Sampling time (Δt)	30 minutes
PHS nominal capacity ($P_{T:G}$ = 2.83 kW and $P_{M:P}$ = 2.83 kW)	4 kWh
Maximum volume (PHS)	100%
Minimum volume (PHS)	5%
PHS nominal capacity ($\eta_{T:G}$ = 84% and $\eta_{M:P}$ = 84%)	70.6%
MHK system rating	3 kW

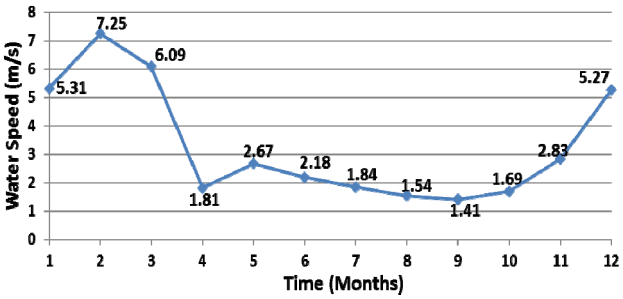


Fig. 2. Monthly average water velocity [17]

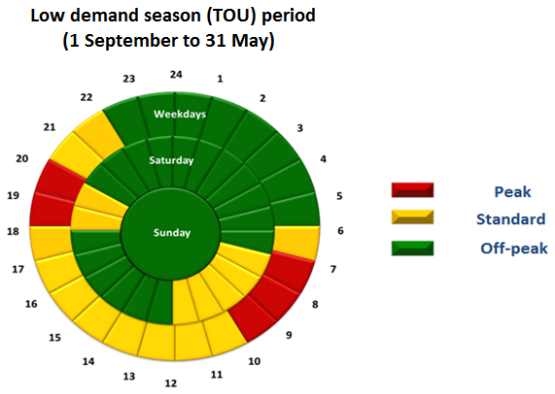


Fig. 3. TOU schedule for low demand season

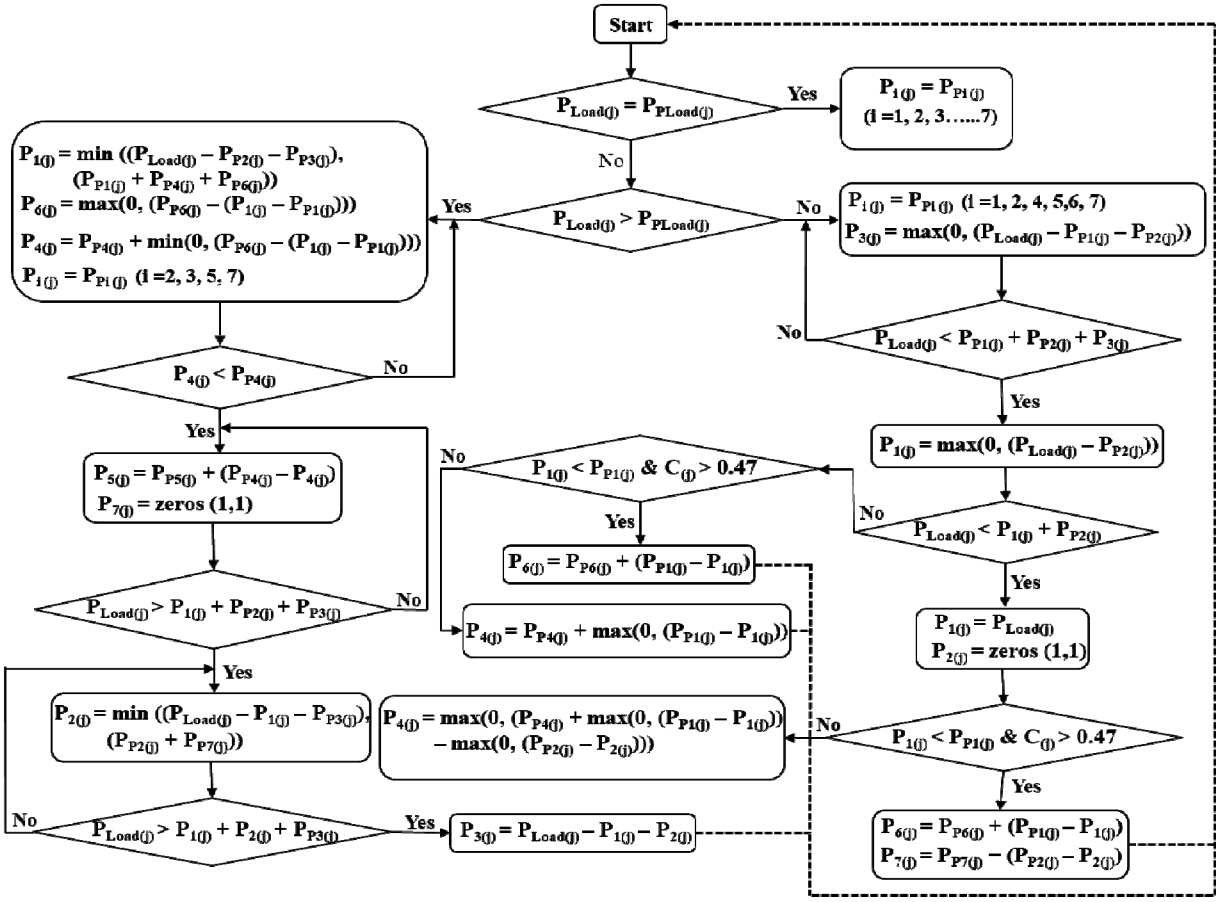


Fig. 4. Flowchart of the developed rule-based control algorithm

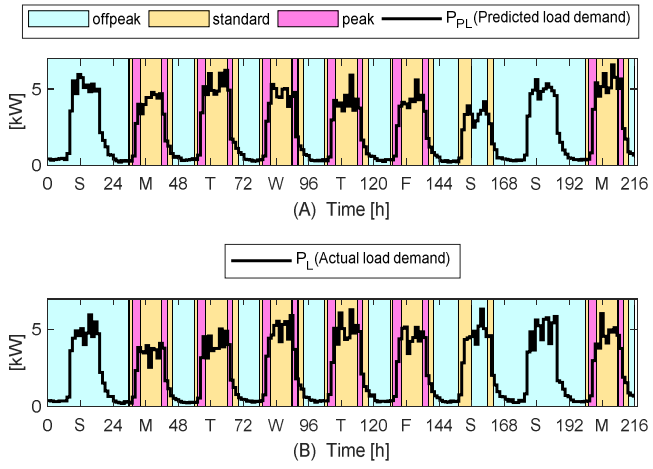


Fig. 5. Commercial load demands (A) predicted (B) actual

VII. Results and Discussion

The simulations have been carried out to analyze the performance of the proposed open-loop optimization approach in combination with the developed rule-based algorithm under load demand uncertainty. The simulations have been carried out for 216 hours (9 consecutive days) at sampling intervals of 0.5 hour, applying the TOU tariff scheme. During energy sales, the FIT is assumed to be 65% of the utility's retail electricity price.

Fig. 5 shows both the predicted and actual load demand as used during the simulation. It can be noted that the two load demands do not consume the same hourly power in order to represent the load demand uncertainty. If the two load demands are solely met by the utility grid for the entire 9 days, they incur a baseline grid cost of ZAR243.25 and ZAR247.15, respectively. However, even though the overall baseline grid costs do not differ that much, the difference in hourly demands may result in a disturbance.

Fig. 6 shows that the open-loop optimization approach is able to meet the predicted load demand. This is before the involvement of the actual load demand. Hence, the predicted load demand is optimally met by the combination of the three power sources (green dotted marks) at no shortage. As soon as the actual load demand is introduced, the open-loop approach is unable to adequately meet the load demand throughout each day, as shown in Fig. 7. This happens during the days whereby the actual load is greater than the predicted load demand, most especially on Wednesday, Thursday and Saturday. As soon as the actual load demand is less than the predicted one, the open-loop approach supply more than the demanded power to the actual load demand.

After the introduction of the developed rule-based control algorithm, the open-loop approach is now able to meet the actual load demand adequately, as shown in Fig. 8. This is achieved by firstly prioritizing to increase the MHK-to-load power when the actual load demand is higher than the predicted one as shown in Fig. 10.

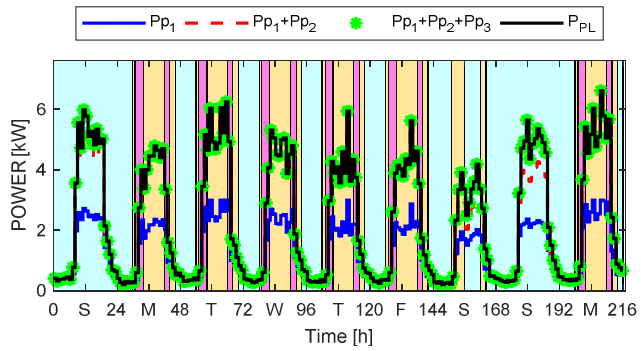


Fig. 6. Open-loop optimal power flow (before load demand uncertainty)

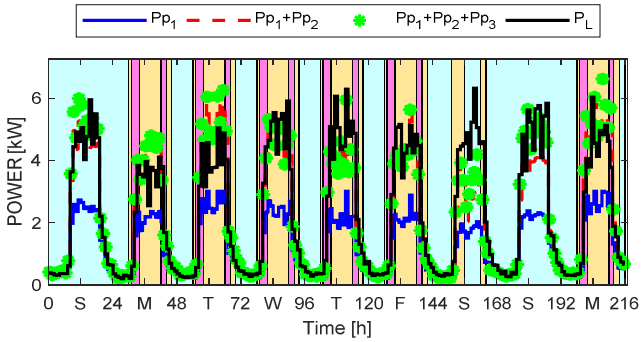


Fig. 7. Open-loop optimal power flow (during load demand uncertainty)

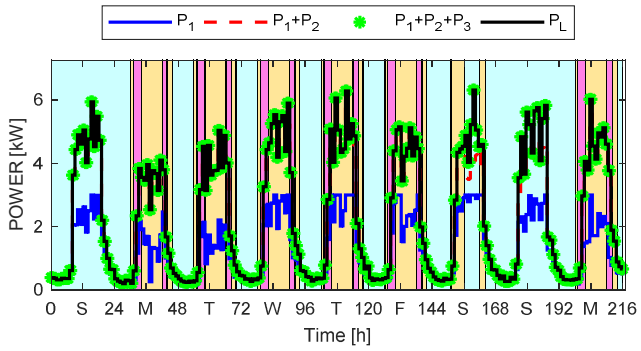


Fig. 8. Rule-based power flow (during load demand uncertainty)

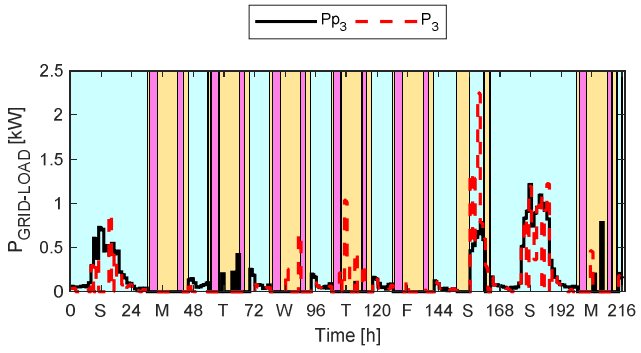


Fig. 9. Grid-to-load power flow

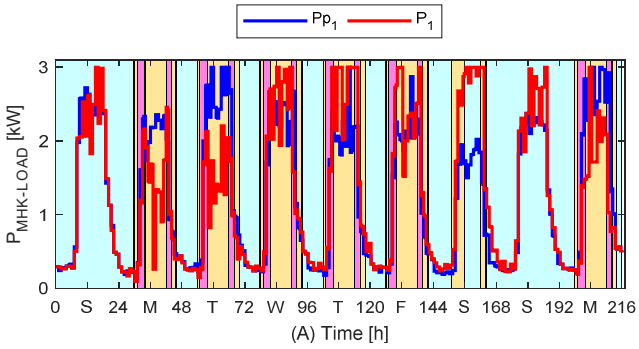


Fig. 10. MHK-to-load power flow

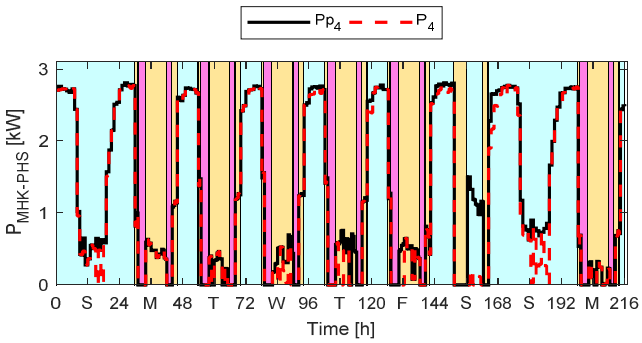


Fig. 11. MHK-to-PHS power flow

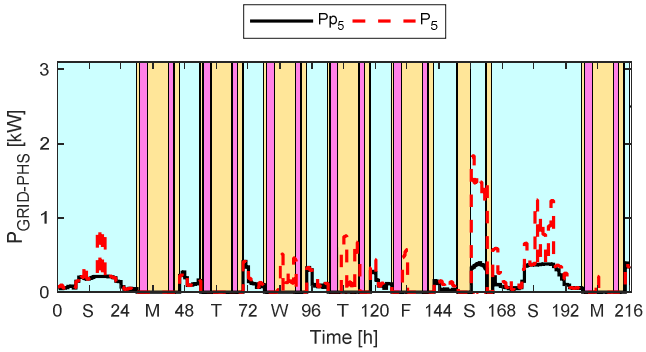


Fig. 12. Grid-to-PHS power flow

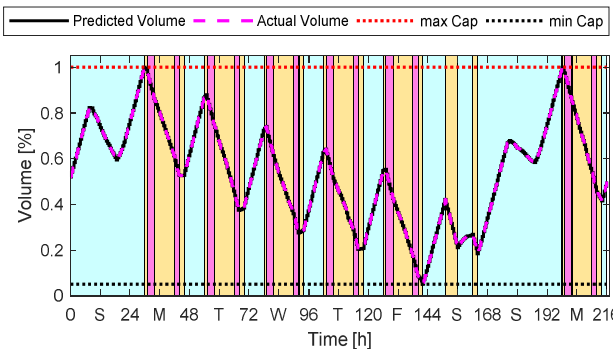


Fig. 13. Upper reservoir state of charge

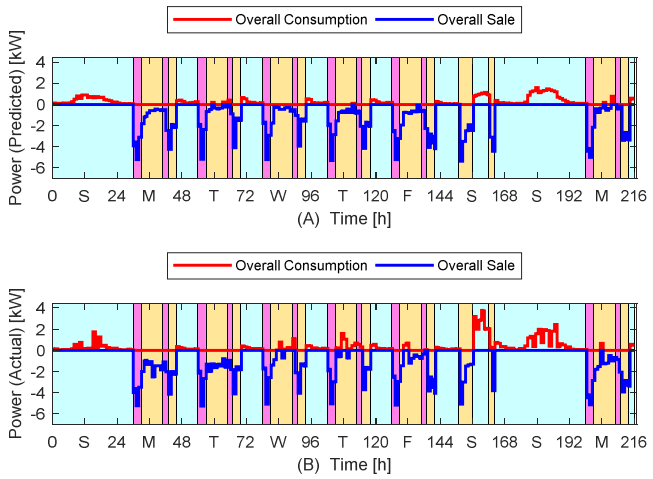


Fig. 14. Overall power consumption and sales

The predicted grid-to-load power is increased only when the RE system cannot solely meet the actual load demand as shown in Fig. 9. As soon as there is an excessive power supplied during open-loop approach, the predicted grid-to-load power is reduced to zero, before reducing the RE power.

When the predicted MHK-to-load power has been increased up to its rated maximum power of 3 kW to supply the actual load, this results in a decrease in MHK-to-PHS power, as shown in Fig. 11. This will affect the level of the upper reservoir to decrease. To maintain the upper reservoir level, the rule-based algorithm allows the grid to supply the difference (P_{P4} - P_4), to the PHS system, as shown in Fig. 12. Therefore, the actual upper reservoir level is maintained to be equal to the predicted level, as shown in Fig. 13.

Fig. 14 shows that the model does not allow simultaneous buying and selling. In order to maximize the sales revenue, most of the energy sales takes place during the expensive peak periods. Additionally, most of the grid energy consumption takes place during cheap off-peak periods. Whenever the predicted load demand is greater than the actual demand, the predicted grid consumption increases to supplement the deficit. Conversely, the predicted energy sales increases. The rule-based control strategy resulted into the overall energy sales revenue and grid consumption cost of ZAR-56.21 and ZAR30.59, respectively.

VIII. Conclusion

The developed rule-based control algorithm proved to convey a guaranteed system reliability to support the open-loop optimization approach. This was achieved without compromising the boundaries of both the power variables and the PHS system. Energy cost saving potential of 110% has been achieved with reference to the baseline grid cost. The algorithm proved to reduce the excessive supplied power, by prioritizing the grid-to-load power reduction. Additionally, whenever there is a power shortage, an increase in RE power is prioritized before permitting an increase in grid-to-load power. Hence, the objective of sustaining a possible minimal grid cost is achievable. For future work, the behavior of the developed algorithm needs to be explored when considering alternative load demand profile types, as well as an alternative disturbance such as the system failure due to maintenance requirement.

References

[1] M.A.H. Baky, M.M. Rahman, and A.S. Islam, "Development of renewable energy sector in Bangladesh: Current status and future potentials", Renewable and Sustainable Energy Reviews, vol. 73, pp.1184-1197, 2017.

[2] A. Mollahosseini, S.A. Hosseini, M. Jabbari, A. Figoli and A. Rahimpour, "Renewable energy management and market in Iran: A holistic review on current state and future demands", Renewable and Sustainable Energy Reviews, vol. 80, pp.774-788, 2017.

[3] A.M. Foley, P.G. Leahy, K. Li, E.J. McKeogh, and A.P. Morrison, "A long-term analysis of pumped hydro storage to firm wind power", Applied Energy, vol. 137, pp.638-648, 2015.

[4] P. Yang, G. Tang and A. Nehorai, "A game-theoretic approach for optimal time-of-use electricity pricing", IEEE Transactions on Power Systems, vol. 28(2), pp.884-892, 2013.

[5] J. Cheng, F. Chu, M. Liu, P. Wu, and W. Xia, "Bi-criteria single-machine batch scheduling with machine on/off switching under time-of-use tariffs", Computers & Industrial Engineering, vol. 112, pp.721-734, 2017.

[6] R. Li, Z. Wang, C. Gu, F. Li and H. Wu, "A novel time-of-use tariff design based on Gaussian Mixture Model", Applied energy, vol. 162, pp.1530-1536, 2016.

[7] A. Kumar and R.P. Saini, "Performance parameters of Savonius type hydrokinetic turbine–A Review", Renewable and Sustainable Energy Reviews, vol. 64, pp.289-310, 2016.

[8] H.J. Vermaak, K. Kusakana and S.P. Koko, "Status of micro-hydrokinetic river technology in rural applications: a review of literature", Renewable and Sustainable Energy Reviews, vol. 29, pp.625-633, 2014.

[9] K. Kusakana, "Techno-economic analysis of off-grid hydrokinetic-based hybrid energy systems for onshore/remote area in South Africa", Energy, vol. 68, pp.947-957, 2014.

[10] H.J. Vermaak, K. Kusakana and S.P. Koko, "Micro-hydrokinetic river system modelling and analysis as compared to wind system for remote rural electrification", Electric Power Systems Research, vol. 126, pp.38-44, 2015.

[11] K. Kusakana, "Feasibility analysis of river off-grid hydrokinetic systems with pumped hydro storage in rural applications",Energy Conversion and Management, vol. 96, pp. 352-362, 2015.

[12] K. Kusakana, "Optimization of the daily operation of a hydrokinetic-diesel hybrid system with pumped hydro storage",Energy Conversion and Management, vol. 106, pp. 901-910, 2015.

[13] T. Ma, H. Yang, L. Lu and J. Peng, "Technical feasibility study on a standalone hybrid solar-wind system with pumped hydro storage for a remote island in Hong Kong", Renewable Energy, vol. 69, pp. 7-15, 2014.

[14] F.A. Canales and A. Beluco, "Modelling pumped hydro storage with the micropower optimization model (HOMER)", Journal of Renewable and Sustainable Energy, vol. 6, 2014.

[15] B.P. Numbi and S.J. Malinga, "Optimal energy cost and economic analysis of a residential grid-interactive solar PV system- case of eThekwini municipality in South Africa", Applied Energy, vol. 186, pp. 28-45, 2017.

[16] A. Kanwar, D.I.H. Rodríguez, J.V. Appen, M. Braun, "A Comparative Study of Optimization-and Rule-Based Control for Microgrid Operation", Proceedings of the Power and Energy Student Summit, Dortmund, Germany, pp. 1–6, 13–14 January 2015.

[17] S.P. Koko, K. Kusakana and H.J. Vermaak, "Grid-interactive microhydrokinetic with pumped-hydro storage: The case study of three South African demand sectors", 2017 International Conference on the Domestic Use of Energy (DUE), Cape Town, South Africa, pp. 83-88, 4-5 April 2017.

[18] S.P. Koko, K. Kusakana and H.J. Vermaak, "Optimal energy management of a grid-connected micro-hydrokinetic with pumped hydro storage system", Journal of Energy Storage, vol. 14, pp. 8-15, 2017.

[19] S.P. Koko, K. Kusakana and H.J. Vermaak, "Optimal power dispatch of a grid-interactive micro-hydrokinetic-pumped hydro storage system", Journal of Energy Storage, vol. 17, pp. 63-72, 2018.

The Presence of Twitter Bots and Cyborgs in the #FeesMustFall Campaign

Yaseen Khan
Department of Information Technology
Durban University of Technology
Durban, South Africa
khanyas786@gmail.com

Surendra Thakur
KZN eSkills CoLab,
Durban University of Technology
Durban, South Africa
thakur@dut.ac.za

***Abstract*—Internet platforms such as Twitter allow cause-related campaigning as well as analysis through the opportunistic classification and aggregation capability provided by the hashtag (#). South African students leveraged Twitter to launch and sustain a campaign now known as the #FeesMustFall campaign. This campaign aimed to lobby government to provide free university education to disadvantaged students. This study examines the #FeesMustFall campaign to determine if automated software robots played a role. The research question was "Did bots and cyborgs play a role in the #FeesMustFall campaign?" 576 823 tweets were harvested, and the data was cleaned by removing duplicate entries. The remaining 490 449 tweets and 90 783 unique users were used to analyze tweet behavior in terms of frequency, volume, content and tweet source. The results show that bots and cyborgs did indeed play a role. This is a significant finding as #FeesMustFall is the first major South African campaign to leverage bots and cyborgs. An important additional finding was the DeBot API revealed 4 bots not found in our harvested tweets while other trait-driven techniques used identified suspicious accounts which revealed two bot or cyborg accounts ranked 1st and 2nd amongst the highest tweeters. This demonstrated a presence of bots during the campaign that assisted in the amplification of the #FeesMustFall hashtag on Twitter.**

Keywords—Social Robots, Software Robots, bots, cyborgs, #FeesMustFall

I. INTRODUCTION

South Africa have recently experienced a unique university student-driven social activists' campaign called #FeesMustFall which dominated local and social media platforms. The #FeesMustFall is a youth movement whose objectives are on reducing university fees with an aim of achieving free education for all [1].The use of Social Media has grown exponentially this past decade nurturing online activism or 'slacktivism'. This use has allowed research particularly on Twitter primarily due to the classification and aggregation capability offered by the hashtag (#). The hashtag allows for a topic to be easily shared worldwide users by enabling others to join the conversation through simply using the same hashtag. This provides the opportunity for easy tracking of topics such as #FeesMustFall. Further, Twitter allows for the development of software robots (bots) or cyborgs to alternatively rebut or amplify a particular point of view pursuant to the views of the originator. #FeesMustFall became desirable to investigate on social media as it is the first major youth driven online campaign in the age of the Fourth Industrial Revolution (4IR) within South Africa generating significant media coverage through headlines bannering '#FeesMustFall'.

The paper next introduces and discusses the applicability of Twitter for analysis of social media campaigns (II) as well as related works on bot detection and deployment (III). The data collection and processing (IV) will then be explained and this is followed by the methodology for identifying automation by users (V), findings and discussion (VI), future works (VII) and the conclusion (VIII).

II. TWITTER AND #FEESMUSTFALL

Twitter proved particularly useful because of the classification and aggregation capability of hashtags. Our particular area of interest with the Twitter platform, was Twitter bots and cyborgs and whether they exerted any influence during the campaign. Twitter bots and cyborgs in the broader spectrum, Social Robots, have impacted campaigns worldwide with the famous being the 2016 United States presidential campaign [2] and the "Arab Spring" [3]. In South Africa, Twitter has been successfully used to assist the rescue of a 'carjacked victim' after the hostage tweeted from the trunk of his hijacked car to his girlfriend, who, in turn retweeted the message. The message which went viral and the victim was saved within 3 hours [4].

Lobbyists, activists and hackers use or deploy bots for their agenda by attempting to manipulate social media users' opinions through automated social engagements. The agenda may be personal, cause-related, political or financial. This paper attempts to highlight the presence of Twitter bots and cyborgs without discussing their textual context and interpreting their intentions which is a social study. Social media platforms contain huge amounts of unstructured opinionated data [5] which require unique analytics to uncover rich information. This is a reason why Twitter and Facebook inter alia gained the interest of researchers.

Further, Twitter was selected as the preferred platform for this task due to the convenient nature of how Twitter data *(tweets)* are categorized simplifying searching, filtering, downloading and analysis. The timeline determined for the collection of Twitter data was from the first mention of the #FeesMustFall, on 21 March 2015, until the 10th April 2017 when the data was received from a professional service provider, Podargos [6], rather than Twitter itself due to high comparative costs [7]. During the data analysis research, intriguing social behaviour exhibited by top tweeting users were found with respect to, inter alia, their frequency, volume and content. Upon further investigation certain users displayed automotive behavior. Examples included tweeting messages with fixed constant intervals (Fig. 1) while another included tweeting multiple times per second for several seconds producing an abnormal burst in tweeting activity (Fig. 2). This paper refers to burst mode tweeting as the posting or retweeting of many tweets within a very short space of time.

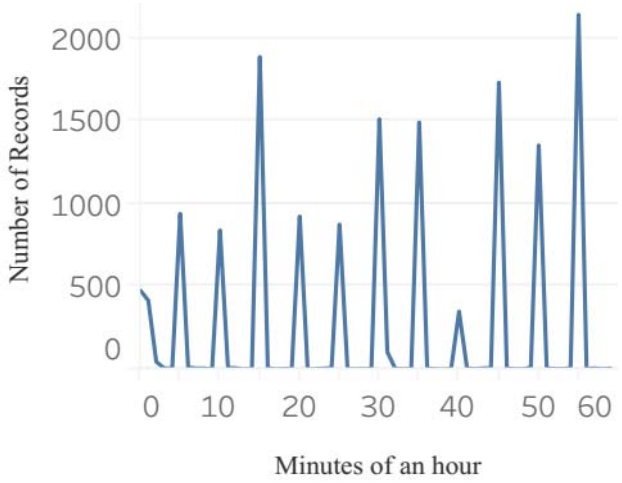


Figure 1. Timeline of Tweets in a one-hour scale for User 1

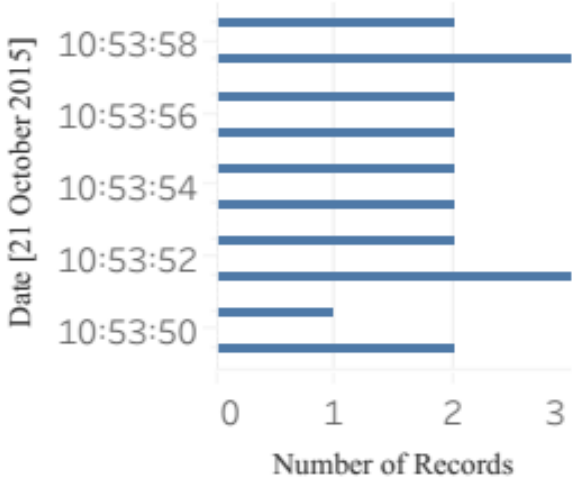


Figure 2. Volume of Tweets within 10 seconds by User 2

Fig. 1 and Fig. 2 depicts users with automating characteristic such as tweeting at specific intervals and multiple tweets per second for several seconds. These characteristics of tweeting is synonymous with Twitter Bots (bots) and Cyborgs [8]. In Twitter a bot is a software that is specially written to troll Social Media and to inter alia, amplify certain tweets, or repeatedly tweet a boutique of tweets either in burst mode or with a certain time-based frequency. This is what piqued the researcher's interest to identify the nature and extent of the impact of Twitter Bots and Cyborgs within our dataset. The 4 methods used to identify automated social behaviour in Twitter from our data, are explained in the Methods section. This paper has significance as it is the first paper in identifying Twitter Bots and Cyborgs within the #FeesMustFall campaign on Twitter.

Bots and cyborgs have several characteristics ranging from simple to highly complex [9]. The simple bot, for example, may tweet exactly after every 15 minutes to a maximum of a predetermined number that slips under Twitter's detection of spam in order to avoid suspension and does not reply to tweets, whereas, the complex bot may, for example, exhibit random tweet patterns and mimic tweets from several different users on a specific topic and can reply to tweets. The detection of bots used in this paper begins with a simple technique with iteratively more complex methods with an aim in the future to build and develop methods to detect complex bots and cyborgs.

III. Bot Deployment and Detection

In the DARPA Twitter bot challenge [10] parties were tasked to develop efficient bot detection approaches on Twitter using any approach they desired. It was found that machine learning techniques on their own were considered insufficient because of the scarcity of training data and that semi-automated processes that combined machine learning were more useful. Popular techniques in detecting bots and cyborgs make use of entropy, spam and account properties components [8], [10], [11], [12], [13]. This is built on the premise and subsequent research that activity by human users differs significantly to bots and cyborgs [8]. A warp correlation finder named DeBot was developed by [14] that detects highly synchronous user accounts over a long period in Twitter and observes that human users cannot achieve such levels of synchronisation over lengthy periods and therefore detected accounts are much more likely to be bots. They have made access to this finder public via an Application programming interface (API) [15] which has been used in this research to identify bots on Twitter during the #FeesMustFall campaign.

As outlined by [8] several features were used in the detection of bots and user accounts which are separated into 3 categories with 'Cyborg' included from the usual human and bot classification. The researchers made use of an entropy-based, a spam detection, an account properties component and a decision maker component to identify bots, cyborgs and humans. Our research, however, focuses on a technique based on volume, frequency and source of tweets to identify automation of tweets. Another publicly-available online service, Botometer (previously BotOrNot), evaluates the degree to which a Twitter account displays similarity to characteristics of known social bots by leveraging at least one thousand features which are grouped into Network, User, Friends, Temporal, Content and Sentiment classes [11].

The conjecture in South Africa was that another campaign related to a political matter was the first evidence of Bot activity. Bots in themselves are amoral – they are programmed to represent the views of their developers. The developers on the other hand may be perceived positive, perceived negative or mercenaries depending on a participant view.

IV. Data Collection and Processing

Tweets were collected from the first mention of the hashtag, #FeesMustFall, in October 2015, until early April 2017. A total of 576 823 tweets were retrieved by purchase from a professional service provider. Limitation exists in accessing the Twitter API to retrieve tweets and the cost of retrieving historical data from Twitter is high [7]. Data went through a cleansing process whereby duplicate, perceived erratic and unintelligible data were removed. This left 490 449 tweets which were stored in a database. Formally each tweet (data point) that was analysed comprised of metadata that included the Tweet text; Date Timestamp; Username; Favorited/Liked; Retweeted; Tweet Source and User Language. Videos and images were excluded from the collection and the study.

V. Methodology for Identifying Automation by Users

To determine the presence of bots or cyborgs 3 different methods have been identified to filter out suspicious bot and cyborg activity for further analyses. An existing bot detecting

technique developed by the DeBot team will complement these methods. Method A detects multiple tweets per instance, Method B detects more than one instances of content duplication of tweets, Method C detects the percentage of tweet sources per user that stem from automated software. Method D uses the API from DeBot. Results from Methods A, B and C will then be sampled and analyzed by the authors to assess whether or not there has indeed been bot or cyborg activity based on frequency, content and volume of tweets.

These methods are as follows:

Method A: Identifying at least 2 tweets by a user on a single timestamp. Assumption that a human user is incapable of tweeting or retweeting more than once on a single timestamp without automated assistance.

Let a tweet = $(tw)_i \ \ where\ i\ , (tw)_i \ \in \ \mathbb{N}$

$$\sum_i (tw)_i = Total\ No\ \ of\ tweets$$

Let a user for a tweet = u_j , $where\ j \ \in \ \mathbb{N}\ and$

$$\sum_j u_j = Total\ No. of\ users$$

Let a Timestamp for a tweet = t_k , $\ where\ k \ \in \ \mathbb{N}\ \ and$

$$\sum_k t_k = Total\ No. of\ Timestamps$$

Each Timestamp is representative of the Gregorian Calendar and Coordinated Universal Time (UTC) and follows the format (*year/month/day hour:minute:second*)

$A\ user, u_j\ is\ assumed\ a\ bot\ or\ cyborg\ if\ at\ Timestamp, t_k$:

$$\sum_i (tw)_{ijk} > 1$$

Method B: Identifying duplication of tweets per user occurring more than once for tweets greater than 29 (i.e. A user that has more than one instance of duplicate tweets)

The number 30 was deliberately chosen by the authors for space as well to create a manageable dataset for further investigation. Due to schedule and trigger mechanisms utilised by bots and cyborgs, assumption is made that duplicate tweets per user appear more frequently in automated assisted accounts than Human accounts. The technique was used on users containing a minimum of 30 tweets in our dataset. Assumption is made that an account with a total number of tweets exceeding 29 and whose duplicate tweets comprises of at least 30% of total tweets in a determined period is considered to be spamming using automation.

Let each unique tweet = $\left(\overline{\overline{\text{tw}}}\right)_{\text{i}}$ where i , $\left(\overline{\overline{\text{tw}}}\right)_{\text{i}} \ \in \ \mathbb{N}$

$$\sum_{\text{i}} \left(\overline{\overline{\text{tw}}}\right)_{\text{i}} = \text{Total No. of unique tweets}$$

A user, u_j is assumed to be a bot or cyborg if :

$$\sum_{\text{i}} (\text{tw})_{\text{ij}} \geq \frac{10 * \sum_{\text{i}} \left(\overline{\overline{\text{tw}}}\right)_{\text{ij}}}{7} \text{ , and } \sum_{\text{i}} (\text{tw})_{\text{ij}} \geq \ 30$$

Method C: Identifying users who have posted most of their tweets from known automated applications. In order to amplify tweet volume and engagement, cyborgs and bots make use of trigger mechanisms that automate tasks on Twitter such as retweeting, posting messages, following and replying to posts. This would involve applications and software to accomplish such tasks. The data was filtered for users who posted at least 30 Tweets within the data and if 70% or more of their tweets come from automated sources as mentioned earlier then assumption is made that these accounts are cyborgs or bots. The parameters for this method have been chosen to increase the likelihood of filtering out automated accounts and create a manageable set of users for analysis. The tweet source was part of the metadata retrieved from the timeline of tweets collected. The assumption adopted is that users whose total tweets are at least 30 and whose Tweet source comprises of at least 70% of a known automated application then these users are considered Bots or Cyborgs. For this paper, "IF This Then That" (IFTTT), Hootsuite, TweetDeck, Tweetcaster and Buffer have been outlined as the set of automated applications for analysis. IFTTT is an applet creator that provides the ability to automate tweeting and retweeting on Twitter [16]. Hootsuite is a Social Media manager that primarily provides the ability to schedule tweets on Twitter [17]. TweetDeck is a Social Media application for managing multiple Twitter accounts and has scheduling of tweets as a feature [18]. Tweetcaster is an application for Twitter users to manage their accounts and has the ability to schedule tweets [19]. Buffer is a Social Media manager aimed at businesses that provides ability to schedule tweets on Twitter amongst other automated features [20].

Let $A =$ {IFTTT; Hootsuite; TweetDeck; Tweetcaster; Buffer}, $where\ A\ is\ a\ set\ of\ automating\ tweet\ sources$

Let the sum of the number of times an automating tweet source appears for the j[th] user be denoted as $(ats)_j$, $where\ (ats)_j \in A,\ j \ \in \ \mathbb{N}$

$A\ user, u_j\ is\ assumed\ to\ be\ a\ bot\ or\ cyborg\ if$:

$$\frac{(ats)_j}{\sum_i (tw)_{ij}} \geq 0.7\ , and\ \ \sum_i (tw)_{ij} > 29$$

Method D: Use DeBot API to identify Twitter Bots with the keyword as '#Feesmustfall'. The DeBot API is made public and is useful in detecting Bot accounts and 'your api key' is given by the DeBot team upon completion of registration on their website [21]. To find out bots for the #FeesMustFall campaign, the Python code used was as follows:

```
import debot
db = debot.DeBot('your_api_key')
db.get_related_bots('#FeesMustFall')
```

VI. Findings and Discussion

Method A reveals a total of 283 Users are either bots or cyborgs while Method B and Method C returned 6 and 135 bot or cyborg prone accounts for further analysis. Method D returned a total of 4 bot accounts by the DeBot API**.** Bot and cyborg detection require varying methods as there are several types of bots and cyborgs that exhibit different character traits, therefore, it is unsurprising that variation of results exists as research in this area is ongoing to combat evolving bots and cyborgs.

A number of users were detected more than once by the varying methods with a few of them belonging to the Top 10 users with the highest number of tweets in the dataset such as User 1 and User 2 in Table 1. Method D revealed users that were not in the original dataset and this could be due to several reasons such as Twitter suspending their accounts when the data was collected, or the accounts being removed by its owners. The comparison of the top 10 users with the highest number of tweets to the users detected from Methods A, B and C was conducted in order to examine the possibility of bots or cyborgs significantly contributing to the discussion on Twitter using the #FeesMustFall in terms of volume.

TABLE 1. TOP 10 USERS WITH THE HIGHEST NUMBER OF TWEETS AND THEIR RESPECTIVE NUMBER OF HASHTAGS, UNIFORM RESOURCE LOCATORS (URLS) AND RETWEETS

User Name	No. of hashtags (#)	No. of URLs in Tweet	Number of Tweets	Retweets
User 1	63817	15362	15403	242
User 2	13665	4111	7018	319
User 3	15684	2215	2318	1025
User 4	533	2294	2258	70
User 5	7355	2221	2193	100
User 6	3388	712	2063	7185
User 7	2206	969	1739	28974
User 8	1561	1096	1053	12034
User 9	1297	298	1041	6622
User 10	259	366	949	4171

In Table 1, User 1 and User 2 were ranked 1st and 2nd respectively amongst the highest tweeters and also appears amongst the users found from using Methods A, B and C. This prompted deeper analyses into the behaviour of these users in terms of frequency, content and volume of tweets. Fig. 1 depicts the pattern of all the tweets from User 1 within a 60-minute scale and it can be notably seen that the user generally tweets in 5 minute intervals which is consistent with a simple bot or cyborg account that is programmed to schedule posting of tweets. Table 1 reflects that on average a tweet from User 1 will comprise of approximately 4 hashtags, 1 URL and not be retweeted. Combining what is found for User 1 in Fig. 1 and Table 1, it can therefore be concluded that User 1 is a bot or cyborg. Fig. 2 displays a snapshot of the tweeting behaviour for User 2 within a 10 second timeframe. It can be seen that there are multiple tweets per second and a total of 21 tweets within 10 seconds which is a prime indicator of autonomous behaviour. User 1 and User 2 are therefore deemed as bots or cyborgs from the 90783 unique users and contributed to approximately 22413 (4.57%) of the 490449 tweets. This is a significant finding about the activity during the #FeesMustFall posts on Twitter as it suggests that social media analysts who aim to get an overview of public sentiment on a specific topic such as this may be blurred into inaccurate interpretations without considering the presence of bots and cyborgs as well as their perceived influence.

The aim of the study was to identify the presence of bots and cyborgs during the #FeesMustFall campaign therefore it is sufficient that at least one bot or cyborg account be found within the dataset and not all of them. In addition, the methods used to filter users into possible suspicious bot or cyborg accounts were based on the general characteristics of simple bot or cyborg accounts [8]. These methods were not tested for performance as this study does not focus on developing bot or cyborg detection techniques.

A. Limitations of the Study

Data collected at one point in time is not necessarily the exact same as data collected at a different point on the same topic at a different time due to the possibility of tweets and accounts being deleted, suspended, removed, changed from public to private and meta data being affected due to changes by users. The methods used does not reflect all bot or cyborg accounts and behaviours in the data.

B. Ethics of the study

The actual names of the users were purposely masked to avoid unnecessary ethical issues that might arise from revealing them. Also, the actual content of the tweets from the identified bot or cyborg accounts were not depicted to prevent tracking of these users as some of them may still be active.

VII. FUTURE WORKS

Bot and cyborg identification is a complex area which requires many techniques and algorithms to produce efficient methods to detect and filter out for social media analytics. Some bot writers will write software to strategically evade detection. Indeed, as [14] suggest humans do not exhibit synchronous behaviour for long periods of time on platforms such as Twitter. The question for future research then is on the characteristic's *length of time*: how long is long, do different demographics exhibit different types of engagement? Create a real-time bot detector for Twitter using more advanced features including timeline and volume of tweets. Can an account be actively defended against bots?

A triangulation of real-world events with bot and cyborg activity on social networks during cyber physical campaigns may assist in identifying the influential bots and cyborgs and by analysing their corresponding sentiment and content researchers could narrow down and predict their intentions. It is important because it shows that not all social media campaigns are entirely human-driven. Identifying the influence of bots may mitigate complex challenges and assist antagonists to reach consensus more quickly.

VIII. CONCLUSION

This paper presents the detection of bots and cyborgs associated with the popular campaign, #FeesMustFall, on Social Media. Bots, may internationally not be a new phenomenon, however, this occurrence is a reasonably new phenomenon for South Africa. By using basic methods to filter out and analyse Twitter accounts for cyborg and bot activity, the researchers have concluded that there were indeed bot and/or cyborg activity during the #FeesMustFall campaign on Twitter. Some of which have had a significant impact on Twitter in terms of amplifying the #FeesMustFall hashtag during the campaign. The important contribution of this study is the detection of bots and cyborgs in #FeesMustFall. This is a unique finding to South Africa where there are no known papers or articles revealing such activity during campaigns of such nature prior to this. The presence of bots and cyborgs during campaigns such as the #FeesMustFall poses concerns for governments and relevant stakeholders where malicious or other types of bots can be created in attempts to influence public opinion which can affect elections and campaigns alike.

REFERENCES

[1] S. Booysen, *Fees Must Fall: Student revolt, decolonisation and governance in South Africa*. Johannesburg: Wits University Press, 2016.

[2] A. Bessi and E. Ferrara, "Social bots distort the 2016 U.S. Presidential election online discussion", *First Monday*, vol. 21, no. 11, 2016.

[3] S. Woolley, "Automating power: Social bot interference in global politics", *First Monday*, vol. 21, no. 4, 2016.

[4] R. Millham and S. Thakur, "Social Media and Big Data," in The Human Element of Big Data: Issues, Analytics, and Performance, pp. 179-194, 2016.

[5] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics", *International Journal of Information Management*, vol. 35, no. 2, pp. 137-144, 2015.

[6] "Welcome to podargos data services!", *Podargos.com*, 2018. [Online]. Available: https://www.podargos.com/. [Accessed: 10- Sep- 2018].

[7] "Pricing", *Developer.twitter.com*, 2018. [Online]. Available: https://developer.twitter.com/en/pricing.html. [Accessed: 08- Sep- 2018].

[8] Z. Chu, S. Gianvecchio, H. Wang and S. Jajodia, "Detecting Automation of Twitter Accounts: Are You a Human, Bot, or Cyborg?", *IEEE Transactions on Dependable and Secure Computing*, vol. 9, no. 6, pp. 811-824, 2012.

[9] E. Ferrara, O. Varol, C. Davis, F. Menczer and A. Flammini, "The rise of social bots", *Communications of the ACM*, vol. 59, no. 7, pp. 96-104, 2016.

[10] V. Subrahmanian, A. Azaria, S. Durst, V. Kagan, A. Galstyan, K. Lerman, L. Zhu, E. Ferrara, A. Flammini and F. Menczer, "The DARPA Twitter Bot Challenge", *Computer*, vol. 49, no. 6, pp. 38-46, 2016.

[11] C. Davis, O. Varol, E. Ferrara, A. Flammini and F. Menczer, "BotOrNot: A System to Evaluate Social Bots", *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*, 2016.

[12] C. Zhang and V. Paxson, "Detecting and Analyzing Automated Activity on Twitter", *Passive and Active Measurement*, pp. 102-111, 2011.

[13] A. Wang, "Detecting Spam Bots in Online Social Networking Sites: A Machine Learning Approach", *Lecture Notes in Computer Science*, pp. 335-342, 2010.

[14] N. Chavoshi, H. Hamooni and A. Mueen, "DeBot: Twitter Bot Detection via Warped Correlation", *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 2016.

[15] N. Chavoshi, "nchavoshi/debot_api", *GitHub*, 2018. [Online]. Available: https://github.com/nchavoshi/debot_api. [Accessed: 10- Sep- 2018].

[16] "IFTTT", Ifttt.com, 2018. [Online]. Available: https://ifttt.com/twitter. [Accessed: 10- Sep- 2018].

[17] H. Inc., "Scheduling - Social Media Marketing & Management Dashboard - Hootsuite", *Hootsuite*, 2018. [Online]. Available: https://hootsuite.com/platform/scheduling#. [Accessed: 10- Sep- 2018].

[18] "TweetDeck", *Tweetdeck.twitter.com*, 2018. [Online]. Available: https://tweetdeck.twitter.com/. [Accessed: 10- Sep- 2018].

[19] TweetCaster for Twitter", *Tweetcaster.com*, 2018. [Online]. Available: http://tweetcaster.com/. [Accessed: 10- Sep- 2018].

[20] "Social Media Management Platform | Buffer", *Buffer.com*, 2018. [Online]. Available: https://buffer.com/. [Accessed: 10- Sep- 2018].

[21] "DeBot", *Cs.unm.edu*, 2018. [Online]. Available: https://www.cs.unm.edu/~chavoshi/debot/api.html. [Accessed: 10- Sep- 2018].

Wireless Communication Technologies in Internet of Things: A Critical Evaluation

G. G. K. W. M. S. I. R Karunarathne
Faculty of Information Technology
University of Moratuwa
Moratuwa, Sri Lanka
sajanikarunarathne@gmail.com

K.A.D.T Kulawansa
Dept. of Computational Mathematics
University of Moratuwa
Moratuwa, Sri Lanka
dilinik@uom.lk

M.F.M Firdhous
Dept. of Information Technology
University of Moratuwa
Moratuwa, Sri Lanka
firdhous@uom.lk

***Abstract*—Internet of Things has gained the attention of almost everybody due to its capability of monitoring and controlling the environment. IoT helps making decisions supported by real data collected using large number of ordinary day-to-day devices that have been augmented with intelligence through the installation of sensing, processing and communication capabilities. One of the main and important aspects of any IoT device is its communication capability for transferring and sharing data between other devices. IoT devices mainly use wireless communication for communicating with other devices. The industry and the research community have proposed many communication technologies for IoT systems. In this paper, the authors present the results of an in-depth study carried out on the benefits and limitations of these communication technologies.**

Keywords— Internet of Things, communication technologies, benefits, limitations, smart devices, layered architecture.

I. INTRODUCTION

Internet of Things (IoT) is a recent computing paradigm that envisages to convert everyday ordinary objects into smart objects [1],[2]. IoT has been identified as one of disruptive technologies of the present time that will change the way perceive and understand the world around us and react to its changes. Advances made in ubiquitous and pervasive computing, embedded devices, communication technologies, sensor networks, Internet protocols and web based applications are the common underlying technologies that help make ordinary devices smart ones and hence made IoT possible [3],[4]. Hence, these technologies are commonly known as the enabling technologies of IoT. In this paper, the authors take a critical look at the different wireless communication technologies that are commonly employed in IoT devices with special emphasis on their features, capabilities, strengths and weaknesses.

This paper consists of five main sections. They are namely, Introduction, IoT and communication Technologies, Related Work, Critical Evaluation of Communication Technologies and Conclusions. Section I sets the context of paper by providing an overview to the subject matter discussed in the paper. Section II discusses the IoT in detail along with its application areas and Section III reviews similar work carried out by other researchers and reported in literature. Section IV carries out a critical evaluation of different wireless communication technologies used in IoT with special emphasis on their strengths and weaknesses. Finally Section V concludes the paper by summarizing the findings along with suggestions for future work.

II. INTERNET OF THINGS

The initial idea for Internet of Things was put forward way back in 1999 by researchers attached to the Auto-ID Center at the Massachusetts Institute of Technology [5]. The initial idea envisaged to put intelligence into everyday objects that they can be connected to the Internet for achieving pervasive communication between any real world object to any other object can be realized. The concept of IoT was officially accepted and formalized by the International Telecommunication Union (ITU) at its 2005 World Summit on Information Society in Tunisia by releasing two reports. These reports presented information on key enabling technologies, market opportunities, emerging challenges and implications for the developing world and described IoT as this emerging paradigm will multiply connections creating an entirely new dynamic network of networks [6].

Since the establishment of the ARPANET, the precursor to the present day Internet, in the 1960s, the number of devices connected to it has increased initially slowly then rapidly after the liberalization of the Internet in the late 1980s [7]. Many factors including the possibility of ubiquitous connectivity and the availability of undecillions of IP addresses with IPv6 have aided the evolution of the Internet of Things [8]. It was forecasted that the number of devices connected to the Internet would reach 25 billion in 2020 from 10 billion in 2014 and surpass 100 billion by 2050 [9]. Figure 1 shows the growth of connected devices on the Internet starting from 1950s to 2050 as forecasted by IBM in 2015.

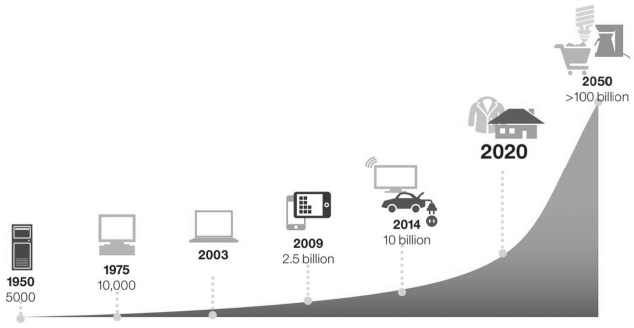


Fig. 1. Growth of connected devices from 1950 to 2050 [8]

Research into IoT has been carried out by many researchers looking at issues from different angles. This has resulted in different architectures proposed by different researchers to suit their requirements. Thus, there is no single architecture that has been accepted by every researcher/user or suitable for every situation [10]. At the early stages of the development of IoT, a three layer architecture consisting of Application, Network and Perception Layers from top to bottom was proposed by Wu et al in [11]. This was a very basic model and many different functions that are today considered under different domains have been combined within a single layer. This five layer architecture was later expanded into a five layer architecture by Tan and Wan in

[12]. The five layer architecture splits the Application Layer of the three layer architecture into three different layers as middleware, application and business layer while maintaining the perception and network layer as they are. Figure 2 shows both the three layered and five layered models as proposed in the literature.

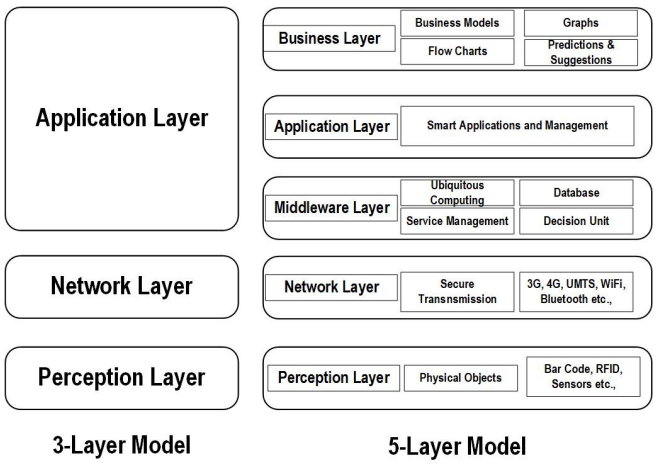


Fig. 2. IoT layered architectures proposed in literature [11],[12]

Tan and Wang have also in [13] have proposed another 5-layered architecture consisting of Edge Technology Layer, Access Gateway Layer, Internet Layer, Middleware Layer and Application Layer from bottom up. Figure 3 shows the architecture proposed in [13]. In this model, both access gateway and Internet layers are responsible for communication and data transmission between devices in the system whereas the same function is included within a single network layer in the models proposed in [12] and [11]. A detailed discussion on the available IoT reference architectures with special emphasis on their current status is given in [14].

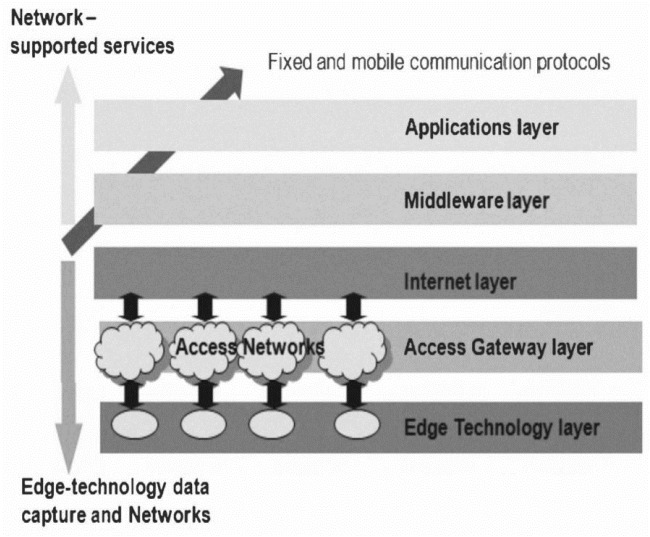


Fig. 3. IoT layered architecture proposed by Tan and Wang [13]

From the above discussion, it can be seen that communication has been given special attention by allocating separate layer(s) within all the reference models put forward so far. This is again common sense as the communication must be given adequate importance to make any distributed system effective and efficient. That is the main reason for the authors of this paper o choose to evaluate the different communication technologies employed in IoT systems.

III. Related Work

This section takes a brief look at other surveys, evaluations and analyses carried out on IoT communication techniques by other researchers. The main objective of this section is to show how this work is different from all the other work reported in the literature.

Longoria et al., in [15] have proposed a novel wireless power transmission system for green smart cities. IoT plays a pivotal role in their proposal as IoT devices are used to manage scheduling, information processing and control of energy supply and demand to households. Though, the main objective of this paper was not to evaluate the communication systems used in IoT, they have carried out an evaluation of them as they will play an important role in the final implementation. The evaluation carried out in this work is very shallow and focused on wireless power transmission rather than communication. Also, all the technologies available are not taken into account for the evaluation as it mainly concentrates on a Cognitive Radio based Software Defined Network (SDN) architecture for managing heterogeneous IoT devices.

Chiariotti et al., have carried out an evaluation of the feasibility of using different communication technologies with IoT for smart cities [16]. The evaluation has taken a critical look at systems like cellular systems, short-range multi-hop technologies and low power wide area (LPWA) technologies with the focus of using them in smart city implementations. The main shortcoming of these communication technologies identified in this study is the lack of possibility for large scale deployments in constrained areas. They further propose that these shortcomings could be overcome by employing different network optimization techniques including cell zooming, implementing HetNets, context-aware content distribution, Cloud Radio Access Network (Cloud-RAN) and network slicing. Though a comprehensive evaluation of communication systems was carried out, the focus was very narrow and confined only to smart city implementations, other domain such as healthcare, disaster mitigation and management, agriculture etc., have not been given any attention.

A proposal for energy efficient data collection in smart cities with the emphasis on how to exploit device to device communication has been presented in [17]. As a part of this proposal, the authors have evaluated device-to-device communication in detail. But the evaluation stops short of evaluating individual communication technique by presenting of a summary of different device-to-device communication technologies.

Low power wide area communication techniques for IoT has been evaluated by Kartakis et al., in [18]. The recent developments in low power technologies have increased their reach to kilometers making them viable solutions for smart city IoT applications. The researchers have identified the limited data carrying capacity of these technologies as the main shortcoming. They have carried a comparison between spread-spectrum technology, ultra-narrow-band technology and conventional narrow-band technology for the implementation of low power wide area networks. This study

is very narrow and focuses only on three technologies with the emphasis on power consumption. other parameters and requirements of wireless technologies are not given any attention in this study.

A comparative study on the merging wireless technologies in the field of IoT has been presented in [19]. In this work, the capabilities of IEEE 802.15.4 technologies, bluetooth low energy, Wi-Fi IEEE 802.11ah and LoRa technologies have been evaluated. This evaluation concentrates only on low power technologies assuming only they can fulfill the needs of IoT. Thus, the evaluation has limited applicability in the real world applications of IoT such as environment monitoring for disaster mitigation.

Mukherjee and Biswas have taken a look at two different wireless technologies, namely Mobile Adhoc Network (MANET), Wireless Sensor Network (WSN) along with Radio Frequency Identifiers for implementing Smart City applications [20]. Authors of this paper have proposed a multi-level architecture for the communication network with different communication technology for each layer. Though, this proposal is comprehensive, there is no justification for selecting a particular technology is provided or advantages and disadvantages of similar technologies for employing at a specific level is evaluated.

Garcia et al., and Yaqoob et al., have taken a look at different wireless technologies for implementing smart city solutions in [21] and [22] respectively. Even in these papers, the evaluation has a very narrow focus targeted only towards smart city applications.

IV. Wireless Communication Technologies used in IoT

Many different communication technologies are available today for using with IoT. Every technology has its own advantages along with their inherent limitations. This combination of benefits and limitations of the technologies results in no single technology suitable for every situation, requirement and application. Hence, it is necessary to weigh each and every technology available in the market against the requirements of the application to select the most appropriate technology suitable for meeting the conditions. The following paragraphs present a critical evaluation of the most commonly used wireless technologies in IoT.

Bluetooth [21-29] and **Bluetooth Low Energy (BLE)** [19],[20],[25],[30],[31] are commonly used communication technologies in IoT for short distance communication. The advantages of these protocols include the ability to handle large amount of data, ability to create ad-hoc connections and the complete standardization of the protocols in the free band. Compared to traditional Bluetooth, BLE has the additional advantage of low power consumption. Lack of built-in security making both the protocols vulnerable to external attacks, low data rate and short range of operations are identified as the main shortcomings of the protocols.

WiFi [19],[22-27],[27],[29],[32-34], **WiMax** [21],[22], [28] and **Wi-Fi HaLow (IEEE 802.11ah)** [19],[25],[28],[31] are another set of protocols that are popular for setting up long distance communication channels. The advantages of these protocols include the ability to transfer both voice and data over long distances at high communication speeds, ability to support large number of client stations with single base station node and built-in support encryption support for enhanced security. On the other hand, the requirement for line of sight transmission and blocking by obstacles and susceptible to signals degradation due to environmental factors, high power consumption and increased latency are identified as the main shortcomings of these protocols.

MiWi and **MiWi P2P** [25],[28],[32] are two protocols that support communication over medium distances at medium data rates. These protocols come with the additional advantage of zero licensing cost. Hence these protocols are suitable for setting up of medium sized IoT networks with not so large number of devices. The proprietary standard and the lack of ability to inter-operate with devices from different vendors have been identified as the main drawbacks of these protocols.

Zigbee [21-32],[34-36] and **Z-Wave** [22],[25],[28],[29], [31],[34] have become popular protocols for smart home implementations due to their advantages of being easy to setup, support for ad-hoc communication with no central control and increased resilience to maintain connectivity with remaining nodes in the face of node failures and low operating cost along low power consumption. The main limitations of Zigbee and Z-Wave include lack of security, low data rate, short range and lack of standardization limiting the freedom to integrate equipment from different vendors.

LoRa and **LoRaWAN** [19],[22],[25],[28],[31],[37],[38] are two protocols that can support large range of communication in an extended star of stars network topology with a large number of client nodes. These protocols also can support variable data rates depending on the application requirement along with the additional ability trade off the speed with range. The main shortcomings of these protocols are the operation in the unlicensed band making it vulnerable to interference from other devices, supporting only point-to-point connections and low bandwidth support.

LTE, **LTE-M** and **LTE-A** [22],[25-28] are standardized protocols for next generation wireless communications. They operate in the licensed band increasing throughput while reducing interference from other devices. Device density in a given area can be increased by reusing the cellular spectrum along with reducing power consumption by using narrow bands for communication. The main downside of these technologies is the high cost of operation arising from cellular data-plans used.

ISA100.11a [25],[31],[36],[39] standard is a for reliable and secure communication at very limited power consumption. This standard also supports bandwidth management in a mesh networking environment for load balancing and better throughput. This standard helps low cost/low complexity deployments with the additional ability to coexist with other wireless technologies. The high cost of devices and low data rates are the main shortcomings of this standard.

Ultra-wideband (UWB) [23-25],[30] and **Wavenis** [29],[30],[34] are two protocols that can be used in very low power applications. These protocols also support reliable communication at very high data rates. The spread spectrum technology used in the physical layer of these protocols provides the additional advantages of immunity from multi-path propagation effects and interference from other devices operating in the same band. The relatively shorter ranges of operation and the difficulty of designing proper antennas are identified as the limitations of these technologies.

Insteon [25],[29],[30],[34] is another protocols that can be successfully used in smart building due to their ability to be used both power-line communication and wireless technologies. Insteon can support long operational ranges with multi-hop transmission in a decentralized architecture avoiding single points of failure. The main shortcoming of Insteon is its proprietary standard limiting the number of devices available in the market along with high power consumption and low data rates.

WirelessHART[25],[28],[31],[32],[36],[39] is a wireless technology that supports self organizing and self healing in a mesh network topology enabling the development of resilient networks. The additional benefits of this wireless standards is its ability to mitigate effects of fading and interference from other signals. This technology supports device authentication and encryption of data for enhanced security of data transmitted. The use of TDMA as the multiple access technology limits the use of free bandwidth reducing the efficiency (throughput) of the network.

Thread [25],[28],[29] comes with native support for IPv6 networking with all its advantages including the support for built-in security, support for different data flows and reduced header functionalities. Thread also supports mesh networking avoiding single points of failure and enabling better scalability supporting a large number of networking devices. The additional benefits of this standard include support for very low power operations and high throughput along with the ability for easy migration from Zigbee, WirelessHART, 6loWPAN etc. Short range and complex nature of the protocol design resulting in limited user friendliness can be considered as the main drawbacks of Thread.

Li-Fi [24],[33],[40],41] is a high speed networking technology that mainly makes use of light for communication. Due to the nature of light used, it comes with the advantages of faster communication and high data transfer rates with low power operations and enhanced security. Since, light is used as the medium of communication along with the requirement of very strict alignment of the transmitter and the receiver for line of sight operations, this technology cannot be employed for outside applications limiting is usefulness. Also the high cost of devices and the requirement for a completely new infrastructure also limits its adoption as a common communication technology by a large number of customers.

EnOcean [25],[28-31] is a communication technology that supports total battery-less operation through energy harvesting from the environment. These devices can harvest energy from multiple sources such as light, motion and pressure. EnOcean's ability work in a battery-less fashion and harvest energy from the environment makes it the most sought after technology for long term deployments of communication devices in remote locations. The other advantages of EnOcean are the supports for high data rates and the ability to work both indoor and outdoor equally well. Limited range and supporting only point-to-point configurations along with nonstandard proprietary technology are considered to be the main limitations of EnOcean.

The above paragraphs presented the results of the investigation carried out on several wireless technologies used in IoT. The main emphasis of the investigation was to find out why such a large number of technologies are available for addressing a single requirement of inter device and device to central node communication. The above analysis clearly shows that each technology tries to identify different communication requirement demanded by different applications and a single technology cannot meet the requirements of all the applications. Hence, every technology has its own advantages along with its limitations. Also, it can be seen that even with such a large number of technologies already available in the market, there is still room for future research as the emerging applications in this field provide more and more challenges with very stringent requirements.

V. Conclusions

In this paper, the authors presented the results of the critical evaluation carried out on the benefits and limitations of different wireless communication technologies used in IoT. Presently there are any wireless communication technologies available in the market and more are continued to be developed by the industry and researchers alike. Every technology has its own advantages and disadvantages. Some technologies are suitable for indoor applications while others can be effectively deployed outdoors. Some technologies perform better when there are only a few peers while others maintain their quality of operations even when large number of peers join in. Hence, it is impossible to identify one single technology that is suitable for all the requirements of every application. As future work, the authors propose to carry out a more in depth analysis with the objective coming out with a comprehensive taxonomy of wireless technologies in IoT that will become the basis of all future studies in this field.

Acknowledgment

This paper is the output of the study carried by the 1st author as part of the compulsory course module “Independent Study” of her undergraduate programme. She likes to acknowledge every person, who helped her to carry out this study successfully. Special thanks to Mr. B.H Sudantha, who was one of her supervisors for this course module.

References

[1] J. E. Ibarra-Esquer, F. F. Gonzalez-Navarro, B. L. Flores-Rios, L. Burtseva, and M. A. Astorga-Vargas, “Tracking the evolution of the Internet of Things concept across different application domains,” Sensors, vol. 17, no. 1379, pp. 1–24, 2017.

[2] M. F. M. Firdhous, B. H. Sudantha, and P. M. Karunaratne, “IoT enabled proactive indoor air quality monitoring system for sustainable health management,” in Second IEEE International Conference on Computing and Communication Technologies, Chennai, India, 2017.

[3] A. Colakovic and M. Hadzialic, “Internet of Things (IoT): A review of enabling technologies, challenges, and open research issues,” Computer Networks, vol. 144, pp. 17–39, 2018.

[4] B. N. Silva, M. Khan, and K. Han, “Internet of Things: A comprehensive review of enabling technologies, architecture, and challenges,” IETE Technical Review, vol. 35, no. 2, pp. 205–220, 2018.

[5] F. Wang, L. Hu, J. Zhou, and K. Zhao, “A survey from the perspective of evolutionary process in the Internet of Things,” International Journal of Distributed Sensor Networks, vol. 2015, pp. 1–9, 2015.

[6] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, “Context aware computing for the Internet of Things: A survey,” IEEE Communications Surveys & Tutorials, vol. 15, no. 1, pp. 414–454, 2014.

[7] A. R. Sfar, E. Natalizio, Y. Challal, and Z. Chtourou, “A roadmap for security challenges in the Internet of Things,” Digital Communications and Networks, vol. 4, no. 2, pp. 118–137, 2018.

[8] IBM Institute for Business Value, “Device democracy: Saving the future of the Internet of Things,” IBM Corporation, Route 100, Somers, NY 10589, USA, Executive Report-Electronics Industry GBE03620-USEN-04, 2015.

[9] A. Haroon, W. Naeem, M. A. Shah, M. Kamran, Y. Asim, and Q. Javaid, “Constraints in the IoT: The world in 2020 and beyond,” International Journal of Advanced Computer Science and Applications, vol. 7, no. 11, pp. 252–271, 2016.

[10] P. Sethi and S. R. Sarangi, “Internet of Things: Architectures, protocols, and applications,” Journal of Electrical and Computer Engineering, vol. 2017, no. 9324035, pp. 1–25, 2017.

[11] M. Wu, T. L. Lu, F. Y. Ling, L. Sun, and H. Y. Du, “Research on the architecture of Internet of Things,” in 3rd International Conference on Advanced Computer Theory and Engineering, Chengdu, China, vol. V5, 2010, pp. 484–487.

[12] R. Khan, S. U. Khan, R. Zaheer, and S. Khan, “Future internet: The Internet of Things architecture, possible applications and key challenges,” in 10th International Conference on Frontiers of Information Technology, Islamabad, Pakistan, 2012.

[13] L. Tan and N. Wang, “Future Internet: The Internet of Things,” in 3rd International Conference on Advanced Computer Theory and Engineering, Chengdu, China, vol. V5, 2010, pp. 376–380.

[14] M. Weyrich and C. Ebert, “Reference architectures for the Internet of Things,” IEEE Software, vol. 13, no. 1, pp. 112–116, 2016.

[15] G. Longoria, F. Akhtar, and L. Shi, “Wireless power transmission in smart cities: The wIshood wireless smart neighborhood,” in 6th International Conference on Smart Cities and Green ICT Systems, Porto, Portugal, 2017, pp. 317–322.

[16] F. Chiariotti, M. Condoluci, T. Mahmoodi, and A. Zanella, “Symbiocity: Smart cities for smarter networks,” Transactions on Emerging Telecommunications Technologies, vol. 29, no. 1, p. e3206, 2017.

[17] A. Orsino, G. Araniti, L. Militano, J. Alonso-Zarate, A. Molinaro, and A. Iera, “Energy efficient iot data collection in smart cities exploiting D2D communications,” Sensors, vol. 16, no. 836, pp. 1–19, 2016.

[18] S. Kartakis, B. D. Choudhary, A. D. Gluhak, L. Lambrinos, and J. A. McCann, “Demystifying low-power wide-area communications for city IoT applications,” in Tenth ACM International Workshop on Wireless Network Testbeds, Experimental Evaluation, and Characterization, New York City, NY, USA, 2016, pp. 2–8.

[19] M. Elkhodr, S. Shahrestani, and H. Cheung, “Emerging wireless technologies in the Internet of Things: A comparative study,” International Journal of Wireless & Mobile Networks, vol. 8, no. 5, pp. 67–82, 2016.

[20] S. Mukherjee and G. P. Biswas, “Networking for IoT and applications using existing communication technology,” Egyptian Informatics Journal, vol. 9, pp. 107–127, 2018.

[21] L. Garca, J. M. Jimenez, M. Taha, and J. Lloret, “Wireless technologies for IoT in smart cities,” Network Protocols and Algorithms, vol. 10, no. 1, pp. 23–66, 2018.

[22] I. Yaqoob, I. A. T. Hashem, Y. Mehmood, A. Gani, S. Mokhtar, and S. Guizani, “Enabling communication technologies for smart cities,” IEEE Communications Magazine, vol. 55, no. 1, pp. 1–19, 2017.

[23] J. S. Lee, Y. W. Su, and C. C. Shen, “A comparative study of wireless protocols: Bluetooth, UWB, ZigBee, and Wi-Fi,” in 33rd Annual Conference of the IEEE Industrial Electronics Society, Taipei, Taiwan, 2007, pp. 46–51.

[24] M. R. Mallick, “A comparative study of wireless protocols with Li-Fi technology: A survey,” in 43rd IRF International Conference, Chennai, India, 2016, pp. 8–12.

[25] M. Pareek and S. Buriya, “A study of link layer protocols in IoT,” International Journal on Future Revolution in Computer Science & Communication Engineering, vol. 4, no. 2, pp. 355–359, 2018.

[26] J. K. Parmar and A. Desai, “IoT: Networking technologies and research challenges,” International Journal of Computer Applications, vol. 154, no. 7, pp. 1–6, 2016.

[27] W. Mengdi, “Wireless communication technologies in Internet of Things (IoT),” MSc Thesis, Faculty of Technology, Communication and Systems Engineering, University of Vaasa, Vaasa, Finland, 2017.

[28] O. Elijah, T. A. Rahman, I. Orikumhi, C. Y. Leow, and M. N. Hindia, “An overview of Internet of Things (IoT) and data analytics in agriculture: Benefits and challenges,” IEEE Internet of Things Journal, pp. 1–17, 2018.

[29] N. M. Turab, “IoT wireless home automation technologies and their relation to specific absorption rate,” Journal of Theoretical and Applied Information Technology, vol. 96, no. 14, pp. 4597–4609, 2018.

[30] H. Kaur and S. Sharma, “A comparative study of wireless technologies: Zigbee, Bluetooth LE, Enocean, Wavenis, Insteon and UWB,” in International Conference on Recent Trends In Computing and Communication Engineering, Hamirpur,India, 2013, pp. 273–276.

[31] T. Salman and R. Jain, “A survey of protocols and standards for Internet of Things,” Advanced Computing and Communications, vol. 1, no. 1, pp. 1–20, 2017.

[32] A. S. Abdul-Qawy, P. P. J, E. Magesh, and T. Srinivasulu, “The Internet of Things (IoT): An overview,” International Journal of Engineering Research and Applications, vol. 5, no. 12, pp. 71–82, 2015.

[33] P. K. Sharma, Y. S. Jeong, and J. H. Park, “EH-HL: Effective communication model by integrated EH-WSN and hybrid LiFi/WiFi for IoT,” IEEE Internet of Things Journal, vol. 5, no. 3, pp. 1719–1726, 2018.

[34] L. Metongnon and R. Sadre, “Fast and efficient probing of heterogeneous IoT networks,” International Journal of Network Management, vol. 28, no. 1, pp. 1–19, 2017.

[35] S. Haseeb, A. H. A. Hashim, O. O. Khalifa, and A. F. Ismail, “Connectivity, interoperability and manageability challenges in Internet of Things,” in International Conference on Electrical and Electronic Engineering, Johor, Malaysia, 2017, pp. 1–7.

[36] D. Christin, P. S. Mogre, and M. Hollick, “Survey on wireless sensor network technologies for industrial automation: The security and quality of service perspectives,” Future Internet, vol. 2, no. 2, pp. 96–125, 2010.

[37] O. Khutsoane, B. Isong, and A. M. Abu-Mahfouz, “IoT devices and applications based on LoRa/LoRaWAN,” in 43rd Annual Conference of the IEEE Industrial Electronics Society, Beijing, China, 2017, pp. 6107–6112.

[38] H. C. Lee and K. H. Ke, “Monitoring of large-area IoT sensors using a Lora wireless mesh network system: Design and evaluation,” IEEE Transactions on Instrumentation and Measurement, vol. 67, no. 9, pp. 2177–2187, 2018.

[39] Z. Li, Z. Chen, J. Zhang, J. Zhu, and N. Xiong, “The evolution of IoT wireless networks for low-rate and real-time applications,” Journal of Internet Technology, vol. 18, no. 1, pp. 175–188, 2017.

[40] S. Akter, R. F. Olanrewaju, T. Islam, and Salma, “LiFi based automated shopping assistance application in IoT,” in 1st International Conference on Big Data and Cloud Computing, Kangar, Perlis, Malaysia, 2017, pp. 1–6.

[41] M. Cedrick, M. Anandraj, and B. J. de Dieu, “How LI-FI will improve the reliability of internet of things: A review,” International Research Journal of Engineering and Technology, vol. 4, no. 4, pp. 2686–2689, 2017.

Human Language Technology Audit 2018: Design Considerations and Methodology

Ilana Wilken
Human Language Technology Research Group
CSIR Meraka Institute
Pretoria, South Africa
iwilken@csir.co.za

Carmen Moors
Human Language Technology Research Group
CSIR Meraka Institute
Pretoria, South Africa
cmoors@csir.co.za

Tebogo Gumede
Human Language Technology Research Group
CSIR Meraka Institute
Pretoria, South Africa
tgumede@csir.co.za

Karen Calteaux
Human Language Technology Research Group
CSIR Meraka Institute
Pretoria, South Africa
kcalteaux@csir.co.za

***Abstract*—Technology audits can play a significant role in surfacing information which can be used by researchers, policy-makers and funders alike to build a country's research and development system of innovation towards increasing its competitiveness, contributing to its economy and bridging the digital divide. In 2016, South Africa established a Centre for Digital Language Resources (SADiLaR) with the aim of supporting a large research infrastructure programme tasked with bringing South African language resources into the digital age. This paper discusses the design considerations and methodology employed to undertake one of the first projects funded by SADiLaR: an updated audit of human language technology resources in South Africa. The paper aims to provide sufficient information to replicate such a technology audit in other environments. The design considerations aim to ensure a pleasant user experience, in order to facilitate as much input as possible. The approach aims to ensure that a sustainable audit tool is developed which can be hosted by SADiLaR in future.**

***Index Terms*—human language technology, technology audit, language resources, text resources, speech resources, digital humanities**

I. Introduction and needs

The establishment of research infrastructure can play a significant role in South Africa's social and economic development, if such infrastructure programmes create opportunities for innovative national research and development. The National Development Plan by the National Planning Commission acknowledges the need for more investment in research and development [1]. The South African Centre for Digital Language Resources (SADiLaR) [2] was recently established as part of the South African Research Infrastructure Roadmap (SARIR) [3]. SADiLaR aims to address the need for access to large corpora of authentic digital data and applicable software tools to enable South African researchers to advance localised research endeavours in the Humanities, Social Sciences, and Information and Communication Technologies in order to address the challenges of unemployment, poverty and inequality [1].

However, researchers, educators, developers, service providers and funders need a roadmap to enable them to decide where to concentrate their efforts in order to give a maximum push to the development of a particular field [4], and to know what is available to enable further technology development and research. Technology audits are an important instrument to provide such a roadmap, with the result that in 2017/8, SADiLaR funded a project to undertake an audit of human language technology (HLT) resources in South Africa. The 2018 HLT Audit aimed to provide updated information on the maturity and availability of HLT resources in the country.

The first-ever HLT audit was conducted by the European Network of Excellence in Human Language Technologies (ELSNET) [5] in 1991, and was based on the idea of a roadmap where information on HLT resources would be collected on a continual basis [6]. The dynamic nature of the ELSNET audit made it suitable for the fast changing nature of the HLT field and therefore suitable to be adapted to similarly gather information on the HLT resources available in South Africa. As a result, the audit of South African HLTs, undertaken by Sharma Grover in 2009 [7], took the ELSNET audit as point of departure.

The sections that follow discuss the methodology, the audit design and approach, a description of the audit instrument development, the execution of the audit and finally a brief analysis of the data and the conclusion.

II. Methodology

The 2018 Audit commenced with a process aimed at identifying and understanding the frameworks available to conduct HLT audits. This investigation uncovered few such frameworks, although a substantial number of references to the Language Resources and Evaluation Conference (LREC) [8]

and the Basic Language Resource Kit (BLARK) [9] initiated by ELSNET [6] were found. The ELSNET approach to their audit was to first conduct a workshop with experts in the field. This was followed by sharing the results via a website and inviting the HLT community to provide inputs. The inputs were then workshopped again, with the concept of a BLARK emerging. This process continued in a cyclical fashion with researchers adding information about their work to the website and the BLARK team updating the information on the website.

In 2009, Sharma Grover [7] adapted the BLARK methodologies described above and undertook an audit of HLT resources in South Africa. Taking the Dutch BLARK [10] as point of departure, Sharma Grover redefined all the HLT components in detail and then produced the first detailed audit on South African HLT resources. The 2009 Audit [11] classified the HLT resources into three categories, namely:

- Data
  - Linguistic data sets or collections (speech or text), in a machine-readable form, used to create, evaluate and improve HLT modules.
  - Includes corpora, lexicas and grammars.
- Modules
  - Basic software units or processes usually required to create HLT applications and products.
  - Includes part-of-speech taggers, sentence tokenisers, language models, acoustic models.
- Applications
  - Categories of different application areas where HLT is used.
  - Includes application domains such as speech input, document production, proofing/authoring tools, and translation.

The data gathered in the 2009 Audit was transferred to the National Department of Arts and Culture's Resource Management Agency (RMA), hosted by the Centre for Text Technology (CTexT) at the North-West University. The RMA subsequently obtained access to many of the resources identified in the 2009 Audit, and made these available via a catalogue (containing downloadable resources), and an index (listing non-downloadable resources).

III. Audit design and approach

A. 2018 Audit design process

The 2018 HLT Audit initially aimed to replicate the 2009 HLT Audit, in order to provide comparable data. A similar process to that followed for the 2009 HLT Audit was thus embarked on. This process entailed the following:

- Familiarising ourselves with the 2009 Audit design process, including:
  - The HLT audit terminology development process
  - The HLT inventory criteria selection process
  - The process for defining the HLT components (and selecting priorities)
  - The HLT audit execution process
  - The HLT inventory gap analysis
- Deciding on the resource categories to be included in the design
- Compiling a list of respondents to be approached to participate in the Audit
- Reviewing the 2009 Audit tool (questionnaire) and determining fit-for-purpose for the 2018 Audit
- Obtaining a thorough understanding of the data analysis techniques used in the 2009 Audit.

1) Defining the structure of the 2018 Audit: In designing the 2018 Audit, HLT experts were consulted in order to assist us to modernise the design. In a workshop with these experts, the component categories which form the basis of the audit were updated; inputs into the audit questionnaire were obtained; and a list of institutions which would be approached to participate in the Audit was compiled.

The workshop attendants were divided into two working groups: one for speech resources and one for text resources. The working groups were tasked with the following:

- Reviewing the 2000 components.
  - Determining which components are still relevant
  - Determining which components need to be changed, added or deleted
- Ensuring that components pertaining to all languages are covered.

The working groups agreed that the Modules and Applications categories are no longer applicable. We therefore only included a Data category and combined the Modules and Applications categories into a Software category. A Model category was added for speech components only. The Data, Model and Software categories were then updated with the resource types which fall into each category, and relevant metadata was added to each component.

Once we had updated the data categories and resource types, we needed to develop definitions for each of the resource types and provide technical descriptions to enable respondents to submit their resources under the correct headings. We nominated a sub-group of experts to assist with this task: three experts for text resources and three for speech resources.

2) Identifying the respondents: Parallel to the process of consulting with the HLT experts on the design of the Audit, we compiled a list of all individuals and institutions involved in HLT research and development in South Africa. This list comprises individuals (contacts) at universities, private companies and research institutions.

B. Audit workflow design

Participating in a technology audit can be a very cumbersome process. If the instrument used to collect the data has not been designed carefully, or is not completely fit-for-purpose, it can lead to a poor user experience and create a barrier to participation. The 2009 Audit employed a Microsoft Excel spreadsheet as the tool with which to collect the data. Navigating through the spreadsheet became cumbersome when large amounts of information needed to be entered. Negative

feedback on the usability and user experience of the 2009 Audit instrument, led us to consider alternatives. We elected to use an online survey tool, instead of a spreadsheet.

In designing the workflow for the 2018 Audit, we studied the 2009 Audit questionnaire and discussed it with the HLT experts at the above-mentioned workshop. Based on these discussions, we designed a new workflow for the 2018 Audit. We defined a number of distinct pages, each containing/requesting information on a specific topic:

- The **Landing page** provides a brief introduction on the 2018 HLT audit, including an overview of how the 2018 Audit will work.
- The **Your Information page** allows users to complete their general information such as name, contact information and affiliation. Users can also choose to be contacted by SADiLaR to have their resource uploaded to the resource catalogue or index.
- The **Resource type page** allows users to select the type of resource that they are uploading, such as text, speech or multimodal.
  - The **Resource type - text selection page**. The user then selects whether their resource is Data or Software. Finally, under either Data or Software, the user may then select the resource type which their resource will be classified as.
  - The **Resource type - speech selection page**. The user then selects whether their resource falls under the Data, Model or Software category. Finally, under Data, Model or Software, the user may then select the resource type which their resource will be classified as.
  - The **Resource type - multimodal selection page**. The user then selects Multimodal corpora.
- The **Required information page** allows the user to complete information on the resource they are uploading. This information includes the name, description and keywords associated with the resource, the language(s) (should the resource be multilingual), the availability, and the cost of the resource.
- The **Technical description page** allows the user to complete further technical information on the resource under the Data, Model and Software pages - this is dependent on the resource type selected earlier in the questionnaire.
- The **Availability page** allows the user to indicate the model of distribution and the license associated with the resource.
- The **Quality page** allows the user to select to complete any protocols, standards and quality assurance methods followed in compiling the resource. Should a user select YES to this question, he/she will be prompted to answer follow-up questions that require detailed information.
- The **Documentation page** allows the user to include a more detailed description of the resource which may not have been covered elsewhere, as well as to upload any other documentation related to the resource.
- The **End page** thanks the user for his/her participation in the Audit and acknowledges the partners in the Audit.

Fig. 1 provides a high-level overview of the flow of the survey.

IV. Audit instrument development

A. Methodology and tool requirements

In selecting an appropriate instrument (tool) for conducting a technology audit, various factors need to be considered. These include cost, functionality and hosting, among other things. We defined the following requirements as a basis for selecting an audit tool:

- Client/user requirements:
  - Online tool (cloud-based or hosted in-house)
  - Attractive to the user (modern look and feel)
  - Clear and easy to use
  - Logical flow
- Functionality:
  - Drop down menus
  - Multiple choice options
  - Yes/No questions
  - Short narrative descriptions possible
  - Document/file upload available
- Technical requirements:
  - Accessible free-of-charge (open platform)
  - Accessible to invited participants (managed participation)
  - Multiple simultaneous inputs possible
  - Ability to store (large) documents (in specific format(s))
  - Ability to export to a database
  - Ability to convert raw data into Microsoft Excel format
- Success criteria:
  - Completeness of information received
  - Scalability
- Outputs:
  - Export raw data to Microsoft Excel format (required)
  - Dashboard with a consolidated view of the audit outcome (optional)
  - Transfer to client website/database (required).

B. Selection of an audit tool

We undertook an Internet search for online questionnaire/survey tools which would suit the needs of the 2018 Audit. We compared different tools, and selected an online tool called LimeSurvey [12].

LimeSurvey is leading open source survey software which is available as Software-as-a-Service (SaaS) or as a self-hosted Community Edition. LimeSurvey is a powerful survey tool which is highly customisable. We opted for the Community Edition, as the solution -

- can be self-hosted and is free of charge;

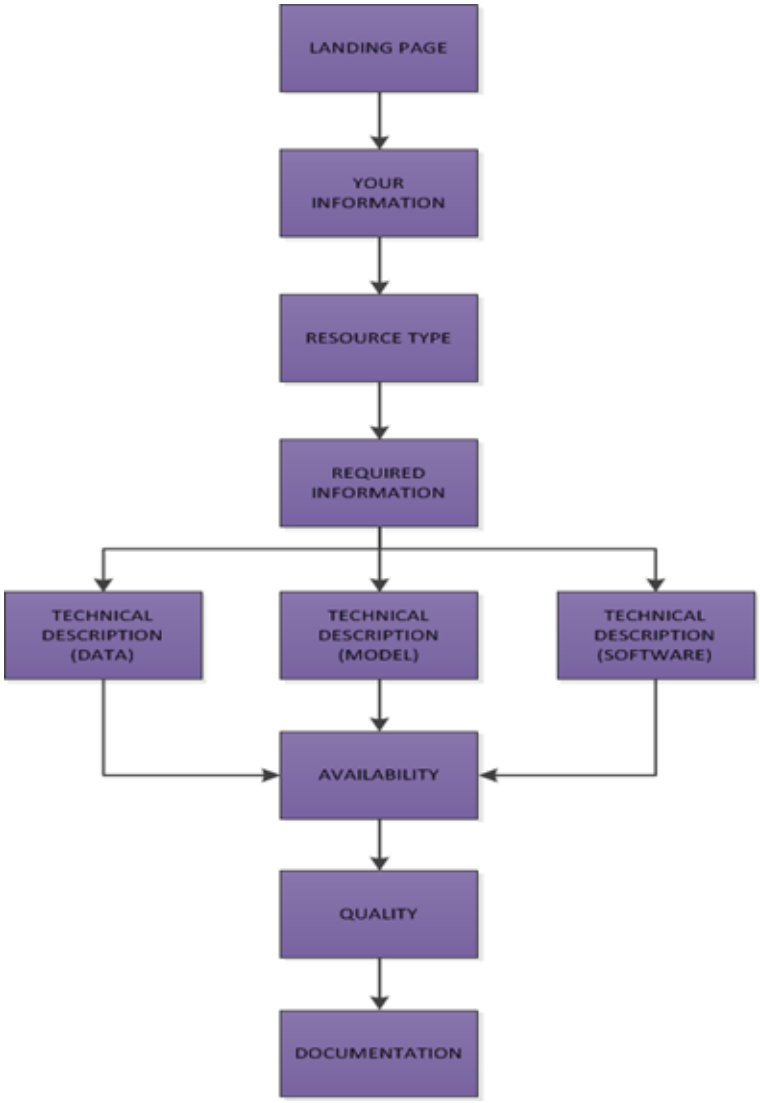


Fig. 1: High-level overview of the 2018 Audit survey

- is easy to set up and customise to the users' needs;
- meets the functionality requirements described above; and
- is accessible using a screen reader.

The user manual and the community forum were then utilised to self-learn the functionalities offered by LimeSurvey.

C. Configuring the audit tool

LimeSurvey offers the functionality of creating a questionnaire using an existing template, or completely from scratch. Since none of the existing templates met the needs of the 2018 Audit, we created a questionnaire.

The properties for every questionnaire created can be changed to suit specific needs. To create a new questionnaire, the following is required:

- Questionnaire title
- Description
- Welcome message
- End message.

There are general settings for each created questionnaire which can be changed as needed. Some of these include:

- Administrator contact details
- How the questions are displayed (question by question vs question group by question group vs all questions on one page)
- Navigation settings (will the user be allowed to navigate backwards or not)
- Displaying the number of questions
- Displaying the progress a user is making
- Access to the questionnaire (open to everyone vs open to anyone who has an access token).

D. Developing the questionnaire

Careful consideration was given to the types of questions to be used for each piece of information required. Usability and user-experience further guided decisions on layout and wording.

The development of the online questionnaire consists of two sections, namely the back-end and the front-end (user interface). The questionnaire workflow was used as the basis for populating the online questionnaire in the back-end of LimeSurvey. Each question was manually created. This entailed:

- Typing the question
- Defining the question type
  - Short text, long text, multiple choice, multiple choice with comments, radio list, radio list with comments, drop down lists, yes/no questions, file upload questions, etc.
- Adding the predetermined answer options (for the multiple choice and radio list type questions)
- Creating conditions for certain questions (for example, "Ask Question 3 if the answer to Question 2 is blue").

E. Beta testing of the audit tool

A beta version of the Audit tool was tested with a small group of beta testers and the feedback was incorporated to the extent possible given the constraints of the online tool. Some of the changes made, based on the feedback received, included:

- Refining/rewording questions
- Changing conditions on certain questions
- Adding an 'other' option to some multiple choice questions
- Adding a list of definitions for the components.

One of the current constraints of the Audit tool, is that it does not allow a user to copy the data from one submitted resource to enable multiple submissions of a similar resource, e.g. where only one or two fields differ across multiple similar resources.

A separate website [13] was also created to provide an easy reference to the list of definitions for the components, as adding all the definitions to the questionnaire would have cluttered the layout and overwhelmed the participants.

F. Providing access to the audit tool

For security purposes, we granted access to the questionnaire by issuing tokens to participants. Each token is valid for a certain number of uses - we set the limit at 100 uses as this was deemed to be sufficient (it is unlikely that one participant would upload more than 100 resources). A unique token was generated per participant and each participant was sent a personalised email containing a link to the questionnaire as well as their unique token.

V. Audit execution

A. Invitation to participate

During the audit design workshop, a decision was made to extend the 2018 Audit to include generic language resources in addition to HLT resources. This was communicated in the email notifying potential participants of the Audit. This email was sent to known members of the HLT community, as well as government departments, the National Lexicography Units of the Pan South African Language Board, publishers, private companies, professional associations, tertiary institutions (we targeted the language, computer science and engineering departments, as well as the language units and requested they disseminate the email to relevant colleagues at the institutions), and the mailing lists of the National HLT Network (NHN) and the Resource Management Agency. The Audit notification email was distributed on 5 and 6 December 2017. The notification email provided background information on the Audit, and requested potential participants to provide us with their contact details should they wish to participate. In addition, the recipients were requested to forward the email to other potential participants within their own networks.

Responses to the notification email generated an automated formal invitation to participate in the Audit. This invitation email contained a link to the online questionnaire (titled "Human Language Technology and Language Resources Audit 2017/8"), the participant's unique token (valid for up to 100 entries), as well as a link to the list of the definitions of the resource components.

B. Responses

The Audit spanned four months, from December 2017 to March 2018. Participants were initially given three months to complete the questionnaire. At the end of month two, follow-up reminder emails were sent out. These were followed by calendar scheduling and phone calls at the end of month three. The latter communication resulted in the extension of the deadline to accommodate additional responses.

A total of 26 completed responses were received. These responses included resources from eight different institutions across South Africa, as well as an institution situated in Germany. Of the 26 responses, 10 were speech-related and 16 were text-related. A total of 76 resources were submitted. An in-depth representation and analysis of the results are presented a paper by Moors, Wilken, Calteaux and Gumede [14]. In the section below, we provide an overview on the process we follow in analysing the Audit data as well as the actual results of the analysis.

VI. Data analysis

A. Introduction and process

The purpose of analysing the data is to determine what the language resource development trends are and to identify if any gaps exist in the availability of resources in specific South African languages. We obtained the actual 2009 HLT Audit data which was uploaded onto the RMA in 2013 and the data that was uploaded from 2014 until the 2018 HLT Audit from the RMA. We were therefore able to cluster the data into three datasets (2009, 2014 and 2018) to be able to compare resource types. We matched the resource types from the 2009 HLT Audit and 2014 RMA data with the resource types modified in the Audit design, as mentioned in section III. From our matching of resource types we were able to compare the availability of a specific resource in a specific

language. We used a graphical representation (stacked column chart) indicating the number of resources submitted per dataset per language (each data set is a different colour). We then tallied the number of text versus speech resources submitted from 2009 until 2018. This provided us with information on resource types which lack resources and languages which have minimal resources. In addition, by subtracting the final number of resources available in 2018 from the resources available in 2009 only, we were able to provide a graphical representation on resource development per language over a period of 10 years.

An example of how we matched and compared the resource types across datasets is as follows: The multilingual lexicon resource type exists in the 2009 HLT Audit, 2014 RMA data and the 2018 Audit. In the 2009 HLT Audit, for English six multilingual lexicons were submitted, in 2014 one more was submitted and in the 2018 Audit another one was submitted. Therefore, a total of eight multilingual resources exist for English from 2009 until 2018. To measure the increase in this resource, we deducted the original six resources from the final number (ten) and converted the difference into a percentage.

B. Results

Based on the comparisons between datasets and calculating the increase in resources, we were able to determine that there is an increase in resource availability for most South African languages. However, languages such as Xitsonga, Tshivenda, Sesotho, siSwati and isiNdebele still remain under-resourced. We were further able to deduce that more text than speech resources are currently available in South Africa.

In addition to the comparison between resource types, we also determined the maturity and accessibility of the resources in all official languages in South Africa. The maturity calculation is based on whether the resource is under development, in its alpha or beta version or released. In terms of maturity, we deduced that speech corpora is the most mature resource type. In terms of accessibility of resources, we used a calculation based on whether the resource is not available/proprietary or closed, if the availability of the resource is undecided, for research or commercial purposes or is openly or freely available. From these calculations, we deduced that text corpora is the most accessible resource type.

A third calculation was done, where the results of the maturity and accessibility calculations were summed for each resource type, in order to get an overview of HLT development in South Africa. Overall, text corpora is the most developed resource type in South Africa, followed closely by speech corpora. Fig. 2 provides an overview of the development of resources in South Africa.

The results obtained from the analysis of the data is able to provide an overview to academics and other interested parties on which resources still need to be developed and in which South African languages. This information is vital for decision-making on resource development investment.

VII. Conclusion

The design and development of the 2018 Audit tool involved extensive research into past and current related audits and methodologies. The experts who participated in this process assisted in creating a simplified and modernised design for collecting information on existing HLT and language resources. The design was implemented in an online tool as method to collect the data. Both the design and the resultant tool can be re-used (with minimal effort) to design future audits (if required) and continually capture HLT resources as these become available.

Future work includes addressing the current challenges with the online tool, particularly the functionality to capture several similar resources with minimal effort. Further work includes implementing a system(s) to ensure that HLT resources (and other language resources) are continually submitted to SADiLaR as these become available. Raising awareness on the benefits of contributing to the body of knowledge and making resources available to others for further research and development, will require focused attention.

Acknowledgments

The authors would like to thank the HLT experts who assisted with formulating and updating the component categories, and providing feedback on the flow of the questionnaire. The same gratitude is extended to all who participated in the 2018 Audit. The 2018 HLT Audit was made possible with the support from the South African Centre for Digital Language Resources (SADiLaR). SADiLaR is a research infrastructure established by the Department of Science and Technology of the South African government as part of the South African Research Infrastructure Roadmap (SARIR).

References

[1] Department of Science and Technology, "South African Research Infrastructure Roadmap. First Edition," Department of Science and Technology, Pretoria, 2016.

[2] South African Centre for Digital Language Resources (SADiLaR), 2018. [Online]. Available: https://www.sadilar.org/

[3] South African Centre for Digital Language Resources, "Research Infrastructure (RI) Proposal for the South African Research Infrastructure Roadmap (SARIR)," 2017.

[4] U. Bross, "Technology audit as a policy instrument to improve innovations and industrial competitiveness in countries in transition," *Innovation: The European Journal of Social Science Research*, vol. 12, no. 3, pp. 397–412, 1999.

[5] European Network of Excellence in Human Language Technologies (ELSNET), 2018. [Online]. Available: http://www.elsnet.org/index_nof.html/

[6] S. Krauwer, "The Basic Language Resource Kit (BLARK) as the first milestone for the language resources roadmap," in *Proceedings of SPECOM*, Moscow, Russia, 27-29 October 2003.

[7] A. Sharma Grover, G. B. van Hyssteen, and M. W. Pretorius, "An HLT profile of the official South African languages," in *Proceedings of the second workshop on African Language Technology (AFLaT 2010)*, Valetta, Malta, 18 May 2010.

[8] International Conference on Language Resources and Evaluation, "LREC Conferences," 2018. [Online]. Available: http://www.lrec-conf.org/

[9] Evaluations and Language resources Distribution Agency, "BLARK: Basic LAnguage Resource Kit," 2018. [Online]. Available: http://www.blark.org/

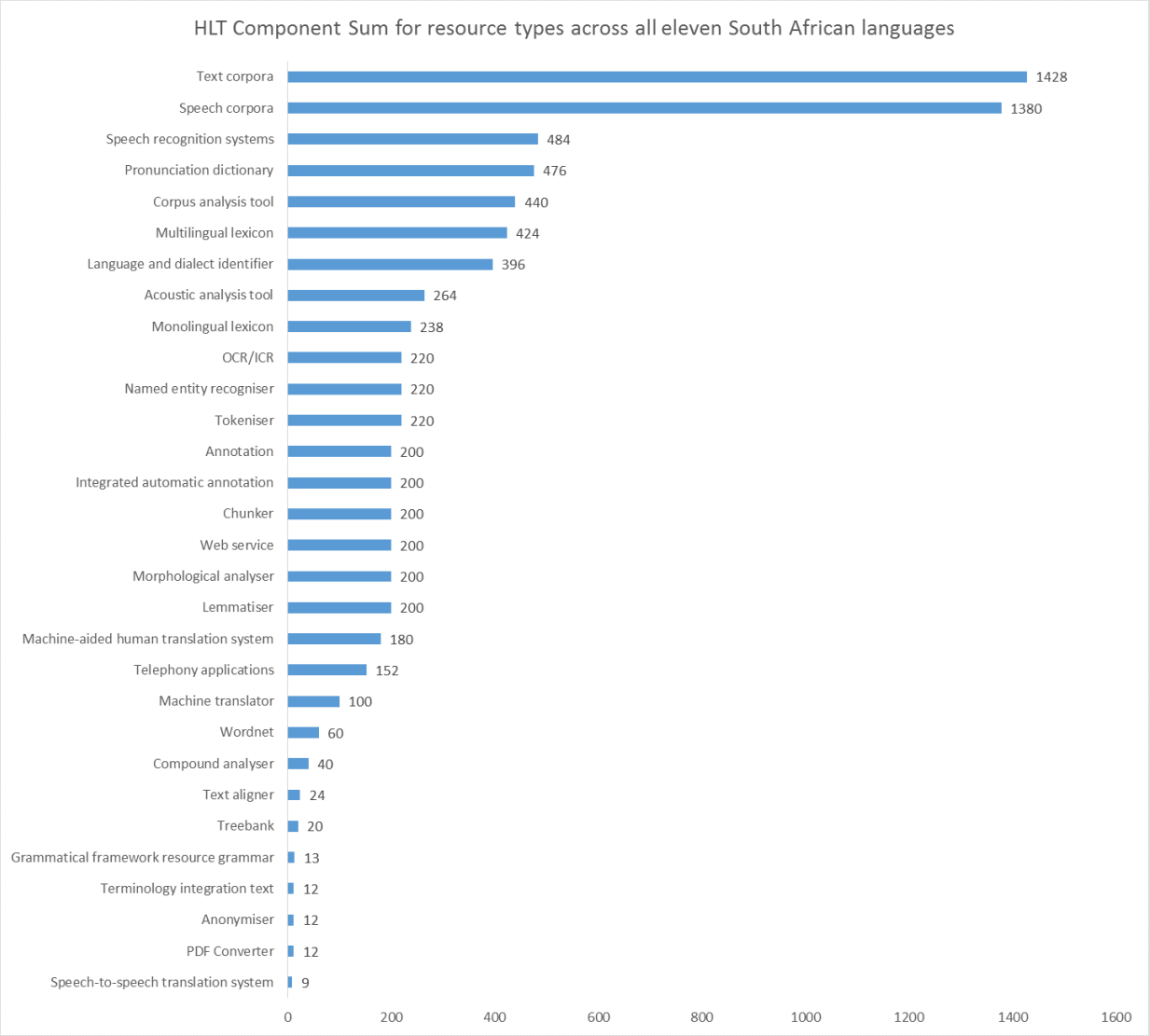


Fig. 2: HLT Component Sums for existing resource types

[10] Helmer Strik, "BLaRK: Basic Language Resource Kit," 2018. [Online]. Available: http://hstrik.ruhosting.nl/blark/

[11] A. Sharma Grover, "Technology Audit: The State of Human Language Technologies R&D in South Africa," Master's thesis, University of Pretoria, South Africa, 2009.

[12] LimeSurvey, "Professional online surveys with limesurvey," 2018. [Online]. Available: https://www.limesurvey.org/

[13] Human Language Technology and Language Resources Audit 2017/8, "List of definitions," 2017. [Online]. Available: https://sites.google.com/view/hlt-audit-defi nitions/home

[14] C. Moors, I. Wilken, K. Calteaux, and T. Gumede, "Human language technology audit 2018: analysing the development trends in resource availability in all South African languages," in *Proceedings of the South African Institute of Computer Scientists and Information Technologists (SAICSIT)*, Port Elizabeth, South Africa, 26-28 September 2018.

Propagation of a jam code signal in the conical-scan seeker processor

Tsholofelo M. Malatji and Cornelius J. Willers
Council for Scientific and Industrial Research
Pretoria, 0001, South Africa
Email: tmalatji@csir.co.za

Maria S. Willers
Denel Dynamics
Centurion, 0046, South Africa

***Abstract*—Infrared guided missiles pose a threat to both military and civilian aircraft, and as such, the development of effective countermeasures against this threat remains vital. The effectiveness of various jamming code signals on the tracking performance of a second generation frequency modulation, conical-scan man portable air defence missile seeker is evaluated in a comprehensive simulation system. The simulation comprises accurate scene geometrical and radiometric models, as well as detailed seeker and kinematic models of the missile. The study investigated the effect a jam signal has on the signal processing circuitry of the conical-scan seeker and compared an effective and non-effective jam signal. It was found that the processor produces a large error if the jam signal oscillates at a frequency that is well suited to the automatic gain control time constant used in the seeker.**

***Keywords*—Electronic countermeasures, jamming, conical-scan seeker, reticle seeker, jamming parameters.**

I. Introduction

The simulation environment is a valuable tool that is often used to model complex systems for various applications. Simulations provide the capability to evaluate scenarios that would be too expensive or impractical to execute in the real world. A simulation allows the user to test many different conditions such as environmental, atmospheric, target and weapon orientations to arrive to some conclusion. This type of analysis would otherwise be impossible using real world field trials. A simulation provides valuable insight into the complex scenario only if the simulation models the physics of the scene accurately. The Optronic scene simulator (OSSIM) is a physics true simulation environment developed jointly by the Council for Scientific and Industrial Research (CSIR) and Denel Dynamics. OSSIM accurately models object signatures in the visible and Infrared (IR) environment, and support missile behavior modeling [1]. OSSIM has been extensively used to develop, test and evaluate IR missile and Countermeasure (CM) systems with high accuracy and favorable outcomes.

This study combines previous work [1]–[3] and reports on the interaction between a jam signal with a conical-scan Frequency modulation (FM) missile seeker while it is guided towards a non-stationary target. This experiment was conducted in an attempt to understand the conditions under which a jam signal would be effective against the FM based conical-scan seeker. The conical-scan seeker was chosen due to the wide proliferation of this second generation Man portable air defence system (MANPADS) [4], [5]. While proliferation of MANPADS has been highly regulated over the years, there remains a high number of first and second generation seekers which are unaccounted for [6]. These devices retain a significantly long shelf life making them a threat even in the current age [7]. This warrants a proper understanding of CM technologies that can be used to counter these threats. While a number of studies have considered the parameters that a jammer signal should contain, it is not clear which jam signals are most effective against the conical-scan seeker [8]–[14]. The present study aims to investigate how a jam signal propagates through the processing circuitry of a conical-scan seeker in order to understand the effect a jam signal has on the seeker.

The paper proceeds with section II which describes the setup of the simulation environment. Section III gives a brief description of how the second generation missile is simulated and how the seeker processes the detector signal. Section IV gives the jam signals that were considered and how they propagate through the seeker. Lastly, Section VI concludes the paper.

II. The simulation environment

A. Geometric modeling

Each object modelled in the OSSIM environment uses a wire frame geometry hull comprising polygons with radiometric properties. The number and size detail of the polygons depend on the level of detail that is required in each section of the object. For example, in the case of an aircraft the engine requires high detail due to the temperature variations that occur in the region surrounding the engine. A large number of small polygons will be used in this region whereas a small number of larger polygons would be used to describe the aircraft fuselage. Fig. 1 shows an example of a helicopter model, but the simulation also has other potential target models available, e.g. fighter and transport aircraft. Each polygon is assigned a material identifier which is associated with specific radiometric properties (Table I). The temperature of each polygon is char-

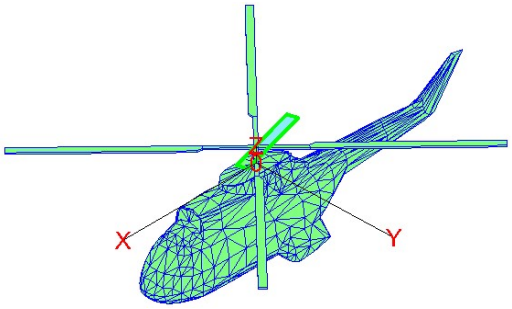


Fig. 1. Wire frame of the helicopter aircraft used in the experiment.

acterized by a temperature equation describing the polygon temperature variation with sun heating, aerodynamic heating and engine setting. Some polygons are modelled with texture properties to account for spatial variation on the object, e.g., a gas plume or CM flare. The polygon properties are determined by combining information from signature measurement campaigns and theory [15] into a coherent, practically and theoretically justifiable and validated model.

B. Radiometric modeling

It is necessary to model the radiometry of the scene as seen by the missile, in order to simulate the missile response during an engagement. The radiation seen by an observer consists of the energy emitted by the source, the reflected energy from all sources and the transmitted energy from all sources [15]. Equation 1 (symbols given in Table I) describes the radiometric contributions modeled in OSSIM (see also [15] for a more detailed description). All these signature contributions are accounted for in the OSSIM simulation environment, hence an accurate signature can be calculated.

$$
\begin{aligned}
L_{\mathcal{S}} = \; & \overbrace{\Delta_\varepsilon \int_0^\infty \varepsilon_{o\lambda}(\theta_v) L_\lambda(T_o) \tau_{a\lambda} \mathcal{S}_\lambda d\lambda}^{\text{thermally emitted } L_{\text{self}}} \\
& + \overbrace{\int_0^\infty L_{\text{path}\lambda} \mathcal{S}_\lambda d\lambda}^{\text{atmospheric path radiance } L_{\text{path}}} \\
& + \overbrace{\int_0^\infty \tau_{o\lambda} \varepsilon_{b\lambda} L_\lambda(T_b) \tau_{abo\lambda} \tau_{a\lambda} \mathcal{S}_\lambda d\lambda}^{\text{transmitted background } L_{\text{trn back}}} \\
& + \overbrace{\Delta_\rho \int_0^\infty \int_{\text{env}} \rho_{o\lambda} \varepsilon_{a\lambda} L_\lambda(T_a) \tau_{ao\lambda} \tau_{a\lambda} \mathcal{S}_\lambda d\Omega d\lambda}^{\text{diffuse reflected ambient background } L_{\text{ref amb}}} \\
& + \overbrace{\Delta_\rho \cos\theta_a \int_0^\infty \int_{\text{sky}} \rho_{o\lambda} L_{\text{sky}\lambda} \tau_{a\lambda} \mathcal{S}_\lambda d\Omega d\lambda}^{\text{diffuse reflected sky } L_{\text{ref sky}}} \\
& + \overbrace{\Delta_\rho \psi \cos\theta_s \int_0^\infty f_r(\theta_s, \theta_v) \varepsilon_{s\lambda} L_\lambda(T_s) \tau_{so\lambda} \tau_{a\lambda} \mathcal{S}_\lambda d\lambda}^{\text{reflected sun } L_{\text{ref sun}}} . \qquad (1)
\end{aligned}
$$

Equation 1 shows that an observer does not only see the radiation emitted by the source of interest, but also the energy that is due to other sources and this contributes to the signature that the observer would see. This means an engagement with a sunlit sky, with no clouds at a specific altitude, is different to a scenario at sunset with a cloudy sky. Reflected sunlight is particularly important when modeling first generation spin-scan Amplitude modulation (AM) seekers, but also has an effect in modeling second generation conical-scan seekers.

C. Atmospheric modeling

The MODerate resolution atmospheric TRANsmission (MODTRAN) software code performs calculation on radiation propagation in the spectral band 100 to 50,000 cm^{-1}. MODTRAN accurately models the signature radiance transmission through the atmosphere and the atmospheric path radiance contributed to the radiance signal. It provides models for various atmospheric conditions encountered in different regions of the world and the effects of different seasons in the year. These models include the effect of aerosols and molecules found in the atmosphere, vertical pressure and temperature profiles, and the effects of sunlight and IR environmental flux scattered through the atmosphere. The OSSIM simulation embeds the full scope of the MODTRAN code through its JMASS interface. The atmospheric properties along any arbitrary sight line (missile to target) can therefore be accounted for.

OSSIM also calculates the background radiance (blue sky or terrain in look-down scenarios) using MODTRAN in order to determine the target contrast signature. This is quite important because the missile responds to target contrast intensity, not absolute intensity.

III. Modeling of the Conical-Scan (FM) Based Missile Seeker

The conical-scan seeker and reticle are modelled and simulated in OSSIM based on the study conducted by Jackman [7]. The proportional navigation guidance law, aerodynamics and flight models were developed by missile experts, providing typical missile guidance and flight behavior for a missile in the MANPADS class. A summary of the OSSIM missile model characteristics are given in Table II.

The optical layout of the seeker is modeled and simulated as shown in Fig. 2. The layout consists of a stationary reticle with rotating optics. The reticle is an optical material which is designed to have transparent and opaque regions which will allow or block the radiation coming through. The reticle used in the simulation is a variation of the wagon wheel reticle (with modified center). The radiation from the scene enters through the seeker dome and is refracted by the primary mirror. The beam then passes through a tilted secondary mirror which provides the optical offset of the nutation circle on the reticle. The reticle performs spatial filtering or chopping of the beam, which is then focussed onto the detector. The rotating optics cause the beam to be nutated on the reticle so that a nutation path is created as the scene is scanned. The nutation process is shown in Fig. 3. As seen in the figure, the target image is nutated on the reticle and not rotated. The image remains upright during the scan and is simply moved across the nutation circle.

TABLE I
TERMINOLOGY DEFINITION FOR EQUATION 1.

Symbol	Meaning
$L_{\mathcal{S}}$	total radiance in the wavelength band $\mathcal{S}$
$L_\lambda(T_s)$	spectral blackbody radiance, sun temperature T_s
$L_\lambda(T_a)$	spectral blackbody radiance, environment temperature T_a
$L_\lambda(T_b)$	spectral blackbody radiance, background temperature T_b
$L_\lambda(T_o)$	spectral blackbody radiance, uniform object temperature T_o
$L_{\text{path}\lambda}$	spectral atmospheric path radiance: emitted & scattered
$L_{\text{sky}\lambda}$	spectral sky radiance: emitted & scattered
$\varepsilon_{s\lambda}$	solar surface's spectral emissivity
$\varepsilon_{a\lambda}$	ambient environment's spectral emissivity
$\varepsilon_{b\lambda}$	background spectral emissivity
$\varepsilon_{o\lambda}(\theta_v)$	object surface directional spectral emissivity
Δ_ε	spatial texture variation in emissivity (unity if no texture)
$\rho_{o\lambda}$	object surface diffuse spectral reflection
Δ_ρ	spatial texture variation in reflectivity (unity if no texture)
$f_r(\theta_s, \theta_v)$	object surface bidirectional reflection distribution function
$\tau_{o\lambda}$	object spectral transmittance
$\tau_{a\lambda}$	object to sensor spectral atmospheric transmittance
$\tau_{abo\lambda}$	background to object spectral atmospheric transmittance
$\tau_{ao\lambda}$	ambient to object spectral atmospheric transmittance
$\tau_{so\lambda}$	sun to object spectral atmospheric transmittance
ψ	$A_{\text{sun}}/(d_{\text{sun}}^2\pi) = 2.17 \times 10^{-5}$
A_{sun}	area of the sun
d_{sun}	distance to the sun
θ_a	angle between the surface normal and the vertical
θ_s	angle between the surface normal and solar incidence
θ_v	angle between the surface normal and the view direction
$\mathcal{S}_\lambda$	measurement instrument spectral response

TABLE II
PARAMETERS FOR SECOND GENERATION MANPADS MODELED IN OSSIM.

Missile diameter	72 mm
Missile mass	10 kg
Missile maximum speed	600 m s^{-1}
Missile roll rate	15 Hz
Lateral acceleration at M1.5	16 g
Drag coefficient (supersonic)	1.3
Field of view (total)	1.9°
Gimbal rate limit	6 °/s
Pre-launch gimbal limit	±45°
Maximum gimbal limit	±60°
Detector waveband	3–5 μm
F number	1.6
Focal length	25 mm

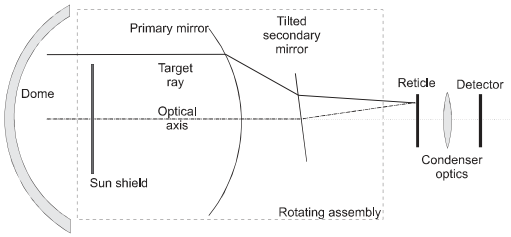


Fig. 2. Optical layout of in modeled seeker.

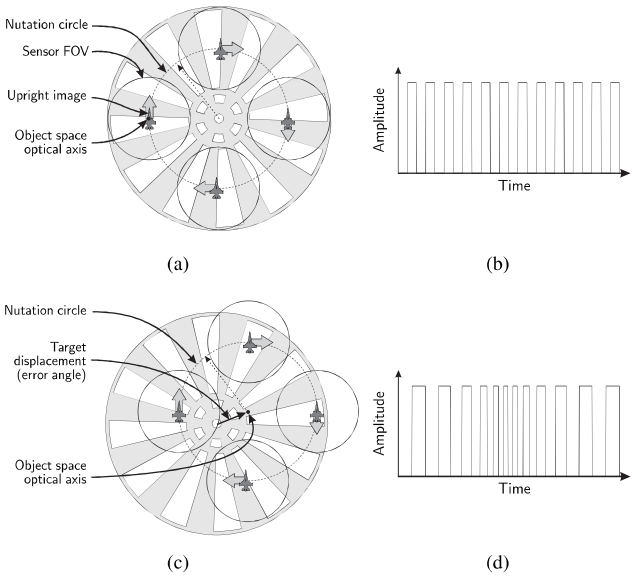


Fig. 3. FM conical-scan reticle (a) with no error, and (b) the associated detector signal, (c) the wagon wheel reticle with a nutation error and (d) the associated detector signal.

The position of the target relative to the optical axis determines whether the nutation circle will be centered on the reticle or not. In the case of the target being on the optical axis (boresight), the nutation circle will be centered on the reticle. This results in a pulse train signal on the detector with no modulation. If the target is off center, or an error exists, the nutation circle will be off center on the reticle resulting in an FM generated detector signal. The modulation depth of detector signal will vary depending on the offset of the nutation circle which indicates the magnitude of the error. A larger

modulation depth in the detector signal indicates a large error, a smaller modulation depth indicates a small error. The phase of the FM signal is used to determine the angular position of the target on the reticle relative to some reference.

The detector signal is processed in OSSIM using the processing block shown in Fig. 4. The detector or sensor signal is passed through a high pass filter which is used to remove the Direct current (DC) component present in the signal. The DC component is generally present due to background noise in the scene and this is removed first. The signal is then normalized using an Automatic gain control (AGC) circuit to ensure consistent envelope amplitude regardless of range or strength of signal.

In the present simulation, the AGC is implemented by using the mean of the previous cycle; each sample value of the current cycle is multiplied by the reciprocal of the mean value of the previous cycle. This signal is multiplied by a gain factor, clipped and normalized to ensure a constant peak signal. Note that any signal frequency modulation will stay intact; general sensor or clutter amplitude interference will be removed. Note that the operation and performance of the AGC has a deciding effect on the effectiveness of the missile's countermeasure performance and hence also on jamming effectiveness. It is imperative that, to evaluate jam code effectiveness against a specific missile, that missile's AGC algorithm must be modelled accurately.

A slope detector is used for demodulation of the FM signal and this is implemented using a Band-pass filter (BPF). The BPF is centered slightly off the carrier frequency and this is done to ensure that the carrier signal lies on the slope of the filter response. Frequencies that are lower than the carrier frequency will be attenuated more than the carrier frequency and frequencies that are higher than the carrier will be amplified more than the carrier signal. Depending on the modulation depth of the FM signal, the slope detector output will produce a relative peak-to-peak output. The amplitude of the slope detector can therefore be used to determine the error of the target signal. The signal is then rectified and filtered to obtain the envelope of the signal. The envelope filter is centered at the nutation frequency of the seeker which is 100 Hz in the simulation. The peak of the envelope signal is compared with the reference signal in the angular phase block. This block determines the angular position of the target relative to the reference. The radial command is obtained by comparing the amplitude of the envelope signal with the reference amplitude. The pitch and yaw signals are then scaled accordingly for accurate torquing of the gimbal.

IV. Jam Signal Processing

A laser jammer was implemented in OSSIM as a directional polygon with a narrow radiation beam. The laser beam was chosen to have a beam width of 1 mrad, this is typical for laser's used in Directed infrared countermeasure system (DIRCM) applications [16]. The laser is mounted on a gimbal and is designed to always direct the beam towards the missile seeker optics. The laser power used in the simulation is 1 W

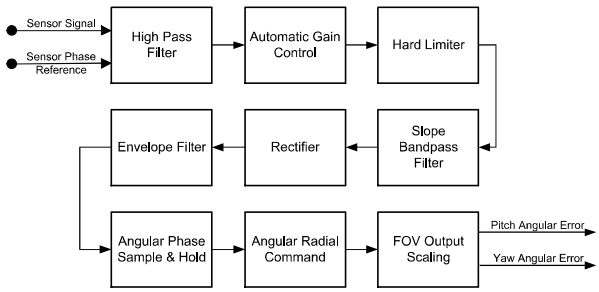
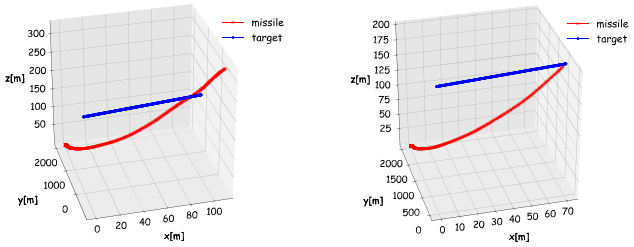


Fig. 4. Demodulation process implemented in OSSIM.



(a) (b)

Fig. 5. Missile and target trajectories for an (a) effective, and a (b) non effective jam signal.

and the laser area is representative of what one would expect of a real DIRCM laser.

Two jam signals were considered for this feasibility investigation. The one operates at a frequency of 25 Hz, while the other at 50 Hz. A Jam-to-signal (J/S) ratio of 10 was used in both cases. The output of the front end high-pass filter (HPF) was used to compute the J/S ratio since the raw detector signal contains a DC offset. The peak to peak value of the target and jam signals were used to compute the J/S ratio.

The relative geometry of the missile and the aircraft flight paths in the simulation world are shown in Fig. 5 for the evaluation scenarios. The main signals produced by the missile seeker processor in the case where the jam signal was effective, are shown in Fig. 6. These signals were recorded over a period of 100 ms for analysis. The detector signal produces a high amplitude during the periods when the jam signal is on and a lower amplitude during the periods when the jam signal is off. The detector signal contains a DC offset as seen in the figure, but this is removed by the front end HPF in the processor. The AGC circuit in this case responds at a time constant of 5 ms. The response of the AGC circuit to the detector signal is shown in Fig. 6(b). The AGC initially adjusts the gain to achieve the desired signal level for the target signal. This is achieved at 5 ms. The jam signal is switched on at the 10 ms mark and due to the gain at which the AGC was operating, the AGC output shoots up as seen in the figure. The AGC immediately responds to reduce the signal to the desired signal level and this is achieved at 17 ms. When the jam signal switches off,

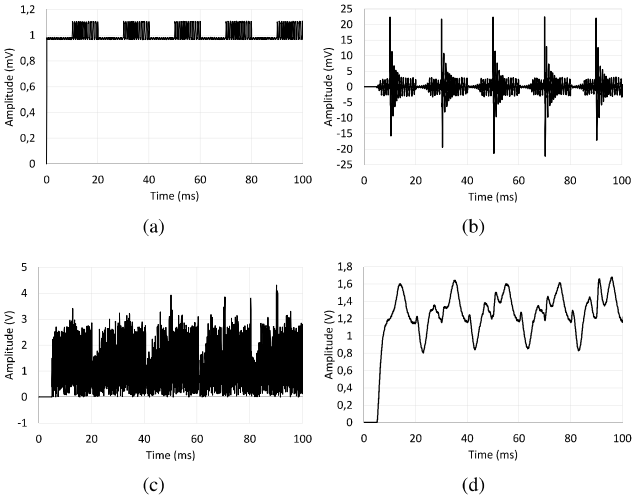


Fig. 6. Output signal of the (a) detector, (b) AGC, (c) rectifier and (d) envelope circuits, for an effective jam code operating at 50 Hz.

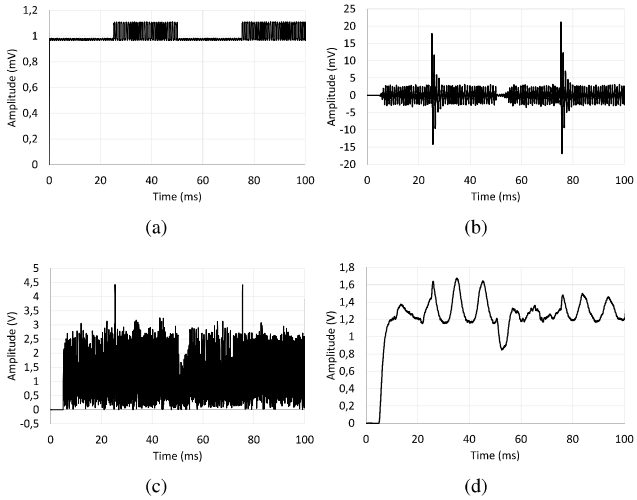


Fig. 7. Output signal of the (a) detector, (b) AGC, (c) rectifier and (d) envelope circuits, for a non effective jam code operating at 25 Hz.

the AGC output drops significantly due to the operational gain. This cycle continues throughout the engagement, making it difficult for the processor to obtain the true target signal level. The signal shown in Fig. 6(c) is the rectifier output signal that has passed though the hard limiter and the slope detector for demodulation. The hard limiter significantly reduces the amplitude variations created by the AGC output. The dips seen in the rectified signal are the dips caused by the jam signal switching off at 20 ms and the multiples thereof. The clipping action of the hard limiter did not successfully remove these. The envelope signal shown in Fig. 6(d), is simply the envelope of the rectified signal. The peak to peak value of the envelope signal is used as an indicator of the error in the signal, and if this is large it indicates a large error. As seen in this case, the envelope signal produces large variations due to the amplitude variations in the AGC output signal and not so much the frequency variations in the signal. This is the masking effect that a jam signal has in the processor. The frequency of this jam signal produces sufficient amplitude variations in the AGC circuit to mask the target signal which is seen minimally. While the frequency variations in the detector signal do produce some AM effect, the effect is minimal when compared to the amplitude variations produced by switching the jammer on and off.

The signals produced by a jam signal that is not effective are shown in Fig. 7. As seen in the figure, the jam signal does produce amplitude variations when the jam signal is switched on and off but due to the AGC time constant, the signal amplitude is adjusted and the target signal is reacquired. The jam signal does not mask the target signal enough to confuse the seeker. The hard limiter significantly reduces the effect of the jam signal and since the number of dips in the signal are few, the effect is reduced. The frequency of this jam signal is slow enough to allow the AGC to reduce the signal and reacquire the target signal. As seen in Fig. 7(b), the processor is able to see the target signal the majority of the time with little disruptions in between. This shows that the jam signal operating at 25 hertz does not adequately mask the target signal. The peak to peak signal produced at the envelope detector is less than that seen in the 50 Hz case. This peak to peak voltage in the envelope signal is used to determine the pitch and yaw errors for steering purposes. With a lower peak to peak variation, the seeker will not be confused by this type of jam signal.

It is important to note that the output of the envelope detector is very much dependant on the signal produced by the AGC and the clipping level of the hard limiter. As seen in Fig. 6, the AGC response time or time constant significantly affects the overall shape of the envelope detector signal. The hard limiter removes the effect of switching the jam signal on but the effect of switching it off is seen in the envelope detector output. As the clipping level decreases, the jam effect is reduced since the integrity of the target signal remains at the low signal level. It can also be argued that a clipping level that is too low can cause more noise to be seen by the processor. This is obviously undesirable. The frequency of the jam signal is quite critical since a frequency too low will result in fewer and less amplitude variations in the envelope detector output resulting in low error. Jam frequencies that are too high though would be filtered out by the envelope detector and the error will not be produced. A combination of the jam signal frequency and the AGC time constant are therefore the critical parameters to consider in achieving successful jamming.

V. Future Work

The results reported here serve to demonstrate the feasibility of using the simulation to develop and test jam codes. The next step would be to evaluate different jam codes versus different AGC algorithms. The suite of jam codes to be considered include various combinations of AM and FM modulations at

different frequencies. The AGC algorithms to be considered includes wide and narrow bandwidth filtered feedback, fast-attack-slow-decay algorithms, and signal statistics algorithms. The basic infrastructure is in place, only the evaluation work now remains.

VI. Conclusion

A physics true simulation environment was used to test the propagation of a jam signal in the conical-scan seeker. It was shown that the simulation environment is well suited for jam code development. A jam signal operating at 50 Hz and another at 25 Hz were used to analyze the difference between an effective jam signal and a non-effective jam signal. It was found that the AGC algorithm and time constant has an effect on the overall error produced in the processor. This shows that the same jam signal can have different effects on the seeker depending on the AGC algorithm used. It was found that the processor produces a large error if the jam signal oscillates at a frequency that is well suited to the AGC time constant used in the seeker. The frequency of the jam signal is therefore critical in defeating the conical-scan seeker.

References

[1] C. J. Willers and M. S. Willers, "Ossim: Optronics system simulator," Jointly published by CSIR and Denel, Tech. Rep. 6700-PG-103400-01/01 RPT, 10504-00004-701, 2016.

[2] C. J. Willers and M. S. Willers, "Simulating the dircm engagement: component and system level performance," in *Technologies for Optical Countermeasures IX*, vol. 8543, 2012, pp. 8543.0M–1 – 8543.0M–16. [Online]. Available: https://doi.org/10.1117/12.974812

[3] T. M. Malatji and W. du Plessis, "Jammer signal effectiveness against a conical-scan missile seeker," in *Africon*, Cape town, South Africa, 18 - 20 September 2017, pp. 591 – 596.

[13] J. R. White, *Aircraft Infrared Principles, Signatures, Threats, and Countermeasures Handbook.* California: Naval Air Warfare Center Weapons Division, 2012.

[4] J. R. Bartak, "Mitigating the MANPADS threat: international agency, US, and Russian efforts," Master's thesis, Naval Postgraduate School, Monterey, USA, 2005.

[5] D. H. Titterton, *Development of Infrared Countermeasure Technology and Systems*. London, UK: Springer, 2006, pp. 635–671.

[6] C. R. Smith, R. Grasso, J. Pledger, and N. Murarka, "Trends in electro-optical electronic warfare," in *Technologies for Optical Countermeasures IX*, vol. 8543, 2012, pp. 8543.02–1 – 8543.02–17. [Online]. Available: https://doi.org/10.1117/12.978652

[7] J. Jackman, "Pre-Emptive Infrared Countermeasures," Ph.D. dissertation, Defence And Security, Cranfield University, 2011.

[8] T. W. Bae, B. I. Kim, Y. C. Kim, and S. H. Ahn, "Jamming effect analysis of infrared reticle seeker for directed infrared countermeasures," *Infrared Phys. Technol.*, vol. 55, no. 5, pp. 431–441, Sep. 2012.

[9] C. J. Tranchita, K. Jakstas, and R. G. Palazzo, "Countermeasure systems," in *The Infrared & Electro-Optical Systems Handbook*, D. H. Pollock, Ed. Raleigh, USA: ERIM and SPIE Press, 2006, vol. 7, ch. 3, pp. 235–285.

[10] T. L. Chang, "The IR missile (spin-scan and con-scan seekers) countermeasures," Master's thesis, Naval Postgraduate School, Montery, USA, 1994.

[11] G. Y. Kim, B. I. Kim, T. W. Bae, Y. C. Kim, S. H. Ahn, and K. I. Sohng, "Implementation of a reticle seeker missile simulator for jamming effect analysis," in *Int. Conf. Image Process. Theory Tools Appl.*, Paris, France, July 2010, pp. 539–542.

[12] F. Qian, J. Guo, J.-F. Shao, and T.-F. Wang, "Study of jamming of the frequency modulation infrared seekers," in *SPIE Int. Symp. Photoelectronic Detection Imaging; Infrared Imaging Appl.*, Beijing, China, 23-25 June 2013, pp. 28–36.

[14] M. C. Sahingil and M. S. Aslan, "The effectiveness of the jammer signal characteristics on conical-scan systems," in *SPIE Conf. Modelling Simulation Defence Syst. Appl.*, Baltimore, USA, 20-24 Apr. 2015, pp. 86–101.

[15] C. J. Willers, *Electro-Optical System Analysis and Design: A Radiometry Perspective*. Bellingham, WA: SPIE Press, 2013, vol. PM236.

[16] B. Molocher, "Countermeasure laser development," in *SPIE Technol. Opt. Countermeasures II; Femtosecond Phenomena II; and Passive Millimetre-Wave and Terahertz Imaging II*, vol. 5989, Bruges, Belgium, 3 Nov. 2005, pp. 16–25.

A Review on Vision Simultaneous Localization and Mapping (VSLAM)

Jabulani K. Makhubela, Tranos Zuva and Olusanya Yinka Agunbiade
Faculty of Applied and Computer Sciences
Dept of Information and Communication, Vaal University of Technology
Vanderbijlpark, South Africa
kingsley.makhubela@gmail.com

***Abstract*—Simultaneous Localization and Mapping (SLAM) has seen a tremendous interest amongst research community in recent years due to its ability to make the robot truly independent in navigation. The capability of an autonomous robot to locate itself within the environment and construct a map at the same time, it is known as Simultaneous Localization and Mapping (SLAM). They are various sensors that are employed in a Simultaneous Localization and Mapping (SLAM) which characterized either as a laser, sonar and vision sensor. Visual Simultaneous Localization and mapping (VSLAM) is when autonomous robot embedded with a vision sensor such as monocular, stereo vision, omnidirectional or Red Green Blue Depth (RGBD) camera to localize and map the environment. Numerous researchers have embarked on the study of Visual Simultaneous Localization and Mapping (VSLAM) with incredible results, however many challenges stills exist. The purpose of this paper is to review the work done by some of the researchers in Visual Simultaneous Localization and Mapping (VSLAM). We conducted a literature survey on several studies and outlined the frameworks, challenges and limitation of these studies. Open issues, challenges and future research in Visual Simultaneous Localization and Mapping (VSLAM) are also discussed.**

Keywords—Navigation; Sensors; Vision; Illumination variance; Simultaneous Localization and mapping (SLAM)

I. INTRODUCTION

Simultaneous Localization and Mapping (SLAM) has captured a great deal of attention within the research community during the past recent years because of it is potential to make robot truly autonomous [1]. Visual Simultaneous localization and Mapping (VSLAM) is when a robot can independently estimate its position within environment and able to draw a map of the same environment, by utilizes vision sensor such as camera, Red Green Blue Depth (RGBD) sensor etc. [2]. Choosing a sensor for autonomous robot such as Laser Finders (LRFs), sonar, acoustic, cameras (monocular, vision stereo or omnidirectional), Red Green Blue Depth (RGBD) Sensor such as Microsoft Kinect and PrimeSense has become a critical part of the SLAM technique [3].

According to [4] vision sensors are utilized in various robotic systems like object recognition, obstacle avoidance, topological global localization. The reason for this is because vision sensor over the other sensors are potable, less expensive, compact, precise, low-priced, non-invasive and pervasive [5].

Vision sensors are able to exact more and viable information both in color and per-pixel about location than any other sensor [6]. Vision sensors are favored because people and animals seem to be navigating effectively in complicated locations using vision as prime sensor [5].

Various researchers have embarked on Visual Simultaneous Localization and Mapping (VSLAM) with exceptional results, however many challenges still exist in Visual Simultaneous Localization and Mapping (VSLAM). This paper will be reviewing the methods, achievements and the limitations on studies done on visual Simultaneous Localization and mapping (VSLAM) by some of these researchers. A reminder of this paper is organized as follows: - section II discusses a review of studies done by some of the researchers. Section III discuss open issues and challenges which still exists on studies reviewed and finally conclusion is drawn in section IV.

II. RECENT RESEARCH ON VSLAM

Visual Simultaneous Localization and Mapping (VSLAM) is when an autonomous robot use service of camera as exteroceptive sensor to navigate, map the location and localize itself. [2]. This section will focus some of the research done by other researchers under a Visual Simultaneous Localization and Mapping (VSLAM), their achievements and limitations they face when implementing their method.

A proposed method by [7] on Stereo Vision Simultaneous Localization and Mapping (VSLAM) for autonomous mobile robot navigation in an indoor location. The objective was to design a system in which an autonomous robot would exclusively utilize a vision sensor for acquiring data and navigating the environment. Their navigation system comprised of navigation and self-localization. The overall navigation hierarchy contained of localization, Perform the Region of Interest (ROI), Region of Interest (ROI) Sub Screening, grid mapping optimal path search and path planning as illustration in Figure 1. The routine activities into their Visual Simultaneous Localization and Mapping (SLAM) navigation system was to achieve 3D depth calculation of the location, scene analysis, optimal path search, real time path planning and motor speed control.

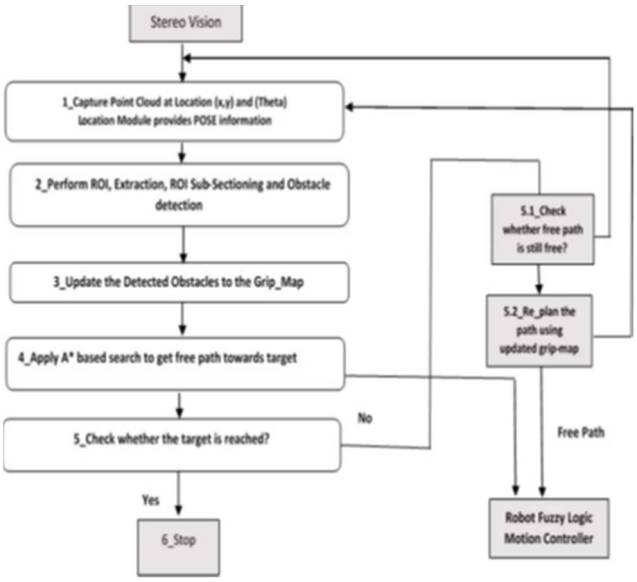


Fig 1. A proposed SLAM Navigation Strategy [7]

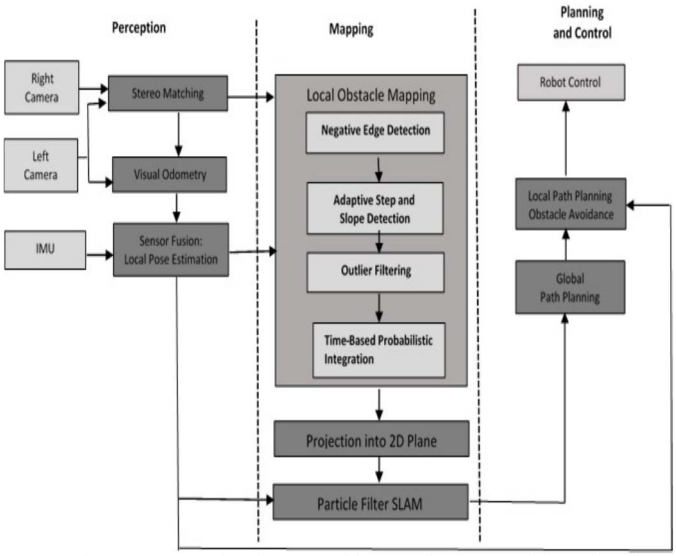


Fig 2. A proposed Software Architecture for Stereo-Vision obstacle mapping [8]

The experimental results showed their proposed system was capable of navigating in a pre-planned paths environment positively. However, the data association on the mobile robot became a challenge due to variations in illumination, specular reflection in environment and inconsistent point clouds which happened because of the variation in viewing angles of the camera.

In the study conducted by [8], for Visual Simultaneous Localization and Mapping (VSLAM) based Indoor/ Outdoor environment obstacle avoidance. The purpose of the study was to create an on-broad stereo-vision based mapping system that would be used for path planning and local obstacle avoidance in a search and rescue mission. Their system design consisted of three layers which was perception, mapping as well planning and control. The perception layer included dense stereo matching, visual odometry computation and sensor fusion. The mapping layer included negative edge detection which assisted in identifying negative slopes and edges in the environment, like cliffs or stair-heads; to further improve on their technique, the adaptive step and slop detection was utilized to convert the depth image into a point-cloud that was associated to the local tangent plane; Outlier Filtering was deployed to reduce the standardization, propagated and stereo correlation error in a 3D positions of the detected obstacles; Time-Based probabilistic Integration was employed to surpass the small aperture angle, and Rao-Blackwellized Particle Filters (RBPF) algorithm it was used to accurately calculate the factorization of jointly probability of the robot route and the environment model into distribution possible robot paths or route. The results of localization and model calculation would be employed by planning and control layer for robot control, local path planning and obstacle avoidance. Figure 2 gives illustration of how the Simultaneous Localization and Mapping (SLAM) system architecture was developed.

The experiments of the proposed system design were performed in an indoor, outdoor and mixed Indoor & outdoor environment. In outdoor set-up the narrow-angle and wide-angle Vision Simultaneous Localization and Mapping (VSLAM) system depicted an above precision of 0.22m mean fault, having an improved deviation of less than 0.08% and 0.06% of the full course meaning that the Vision Simultaneous Localization and Mapping (VSLAM) system was capable to handle inaccuracies, drifting fusion calculations and to rectify the drifts as well. A Vision Simultaneous Localization and Mapping (VSLAM) system also showed favorable progress in closing the large loop in an experimental set-up. The Vision Simultaneous Localization and Mapping (VSLAM) system demonstrated that it could manage with changing light mixed Indoor & Outdoor set-up. However, the system couldn't perform a quantitative evaluation of obstacle map because of lack of ground truth data, the un-textured objects like white walls, regular patterns and reflective surfaces became a problem for the system architecture in an outdoor Environment.

A proposed framework by [9] was presented on Visual Simultaneous Localization and Mapping (VSLAM) in low dynamic workspace environment using a Red Green Blue Depth (RGBD) Sensor, the aim of the framework was to update a map by keeping track on latest changes on environment as the autonomous robot was navigating the location. The proposed framework had two components which were multi-Session Visual Simultaneous Localization and Mapping (VSLAM) and graph management. The multisession Visual Simultaneous Localization and Mapping (VSLAM) module had a graph model with each landmark being a position and each edge being a restriction, by combining previous sessions with the current information and recent session to keep the model of the environment in one global coordinates. The graph management was responsible for updating the graph model and without increasing difficulties employing the out-of-dated scan identification module and redundant scan identification module. Figure 3 gives illustration of the framework.

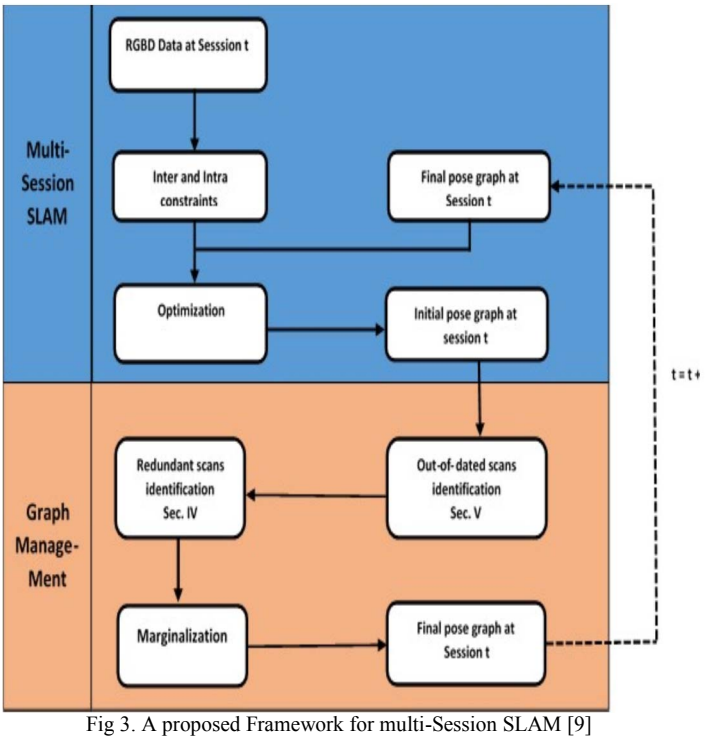


Fig 3. A proposed Framework for multi-Session SLAM [9]

The activities in each timestep was: (1) The multi-session Visual Simultaneous Localization and Mapping (SLAM) module creates model of the environment by employing the Red Green Blue Depth (RGBD) camera information; (2) The out-of-dated scans identification unit was to recognize whether the scans were consistent with previous poses were out of date, if so, the nodes would be removed since they were no longer valuable for creating model of environment and loop closure; (3) The redundant scans identification module would continue removing positions if the number of in-dated nodes was more than a threshold, which was linked to the quantity of the modelling area; (4) The graph was marginalized to reserve the information after the node was removed, establishing an combined restriction for the next session .

The experimental results showed that the proposed framework was able to update the map in a dynamic or static environment without increasing difficulties and at an acceptable error level, however they were setback on the system for out-of-dated scan identification tested in the lab and industrial workspace the small object captured by Red Green Blue Depth (RGBD) sensor with low quality, were considered as noise by the algorithm, two object were regarded as one object and when the object is removed and the new one is added in the position and such change cannot be reflected on the geometric shape.

In reference to a study done by [3] for Robust Visual Simultaneous Localization and Mapping (VSLAM) using planar point's features with the aid of a Red Green Blue Depth (RGBD) sensor. The objective of the study was to propose a method that would improve the accuracy in capturing and recording point's features into the Visual Simultaneous Localization and Mapping (VSLAM). Their proposed method utilized a planar feature to align key frames, the proposed method would start by removing planes points from cloud obtained the Red Green Blue Depth (RGBD) sensor, then attempts improve the quality of the mined planes, then depth values of planes would be modified according to the parameter of plane models and Random Sample Consensus (RANSAC) framework is then employed to sense and compare point features on individual pair of planes. According to [10] Random Sample Consensus (RANSAC) is an overall method for correcting model in the existence of outliers. Their proposed method also contained a graphic end part which executed image process ego-motion calculation and pose optimization part was for executing global pose graph optimization. The illustration of system is given in Figure 4.

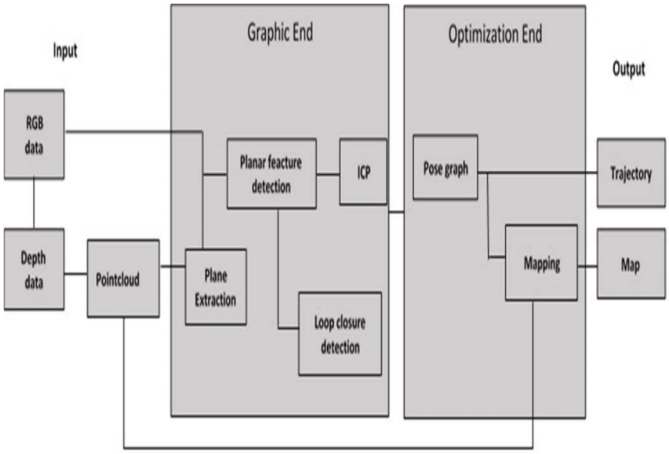


Figure 4. A proposed SLAM Planar Features System [3]

The experimental results of a proposed method on open datasets showed tracking trajectory is fairly accurate, high accuracy in motion of the two fames with less failure rate. However physical experiment results showed repetitive structure on floor, daylight from the ceiling disrupted the matching process and sunshine from window disrupts the illumination conditions.

A proposed algorithm by [11], for robust outdoor stereo Vision Simultaneous Localization and Mapping (VSLAM) for heavy machine rotation sensing. The purpose of the algorithm would be to calculate a mining rope shovel's rotation about its vertical axis. Extended Kalman Filter (EKF) was selected the main algorithm in this study because of its abilities to close loop successfully. [12] defines Extended Kalman Filter as an algorithm that is utilized in a number of nonlinear estimation and machine learning systems as well as calculating the state of a nonlinear dynamic application, calculating parameters for nonlinear system identification when system dynamic and observation models are linear. For every camera frame, based on Extended Kalman Filter (EKF) Simultaneous Localization and Mapping (SLAM), the following processes should take place in their algorithm as demonstrated in Figure 5.

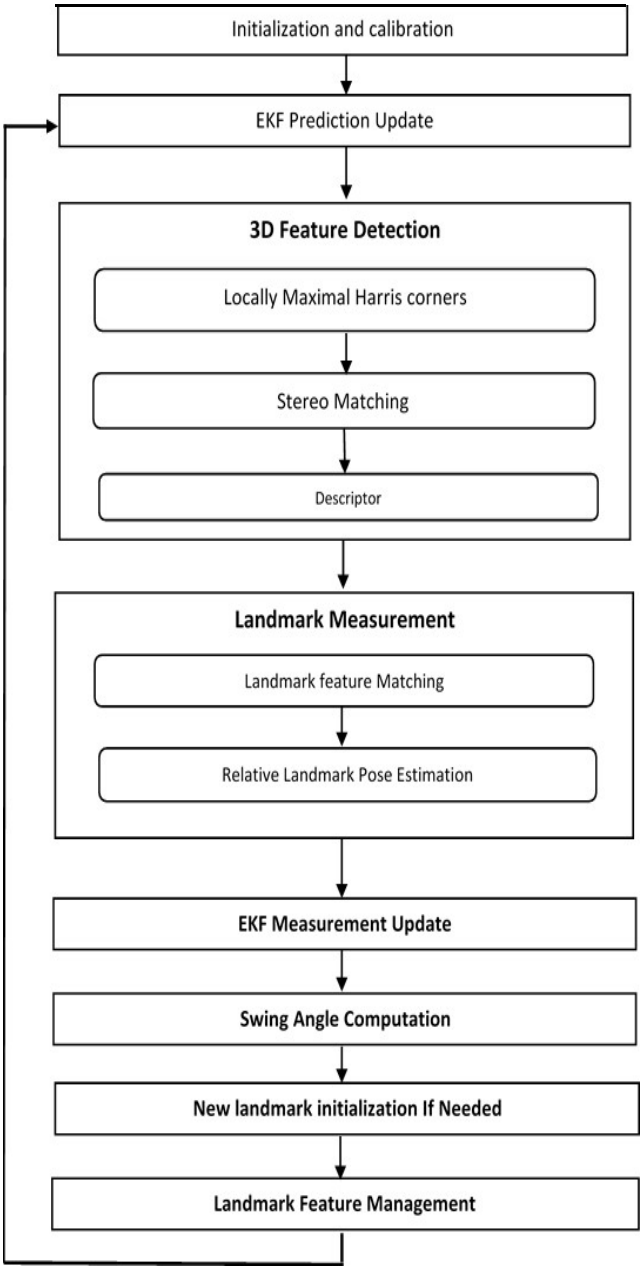


Fig 5. Algorithm flow for robust outdoor stereo vision SLAM [11]

According to the study, the proposed algorithm also employed the two methods to increase Visual Simultaneous Localization and Mapping (VSLAM) effectiveness for outdoor environment. The techniques were Harris corner (Harris & Stephen as cited by [11]), namely locally maximal feature and Feature Cluster landmark. Locally maximal feature chose features more evenly across the image and senses features that consistently in a non-uniformly illuminated scene, Whereas Feature Cluster landmark acquired each landmark in a cluster of 3D feature from a single camera frame, this will permit extremely consistence and robust landmark measurement. As corners were designated based corner threshold within an image. The study also used a percent threshold method to measure the pixel locations of edge that surpasses the corner threshold within an image.

The experiment of a vision algorithm was performed using three separate video sequence, at round 2900, 8000 and 6600 frames respectively. The maximum error for all video sequence settled within 1^{o}. When compared two methods used in a vision algorithm, it was a discovered that the Local maximum method was steadier than Percent Threshold method detecting frame with direct sunlight or shadows illumination.

A framework was proposed by [13] for Collaborative Visual Simultaneous Localization and Mapping (CoSLAM) in a dynamic Environment, the aim of the framework was to address the challenge of dynamic environments by employing numerous cameras into Visual Simultaneous Localization and Mapping (VSLAM) system. The Collaborative Visual Simultaneous Localization and Mapping (CoSLAM) framework would sense and detect feature points of each image from individual camera sensor inputs and provide them into Collaborative Visual Simultaneous Localization and Mapping (CoSLAM) components which are camera pose estimation, map building, point classification and camera grouping which were combined to create a global map. Collaborative Visual Simultaneous Localization and Mapping (CoSLAM) framework would at the same time calculate the position of all camera sensors continually. Figure 6 gives illustration of the framework.

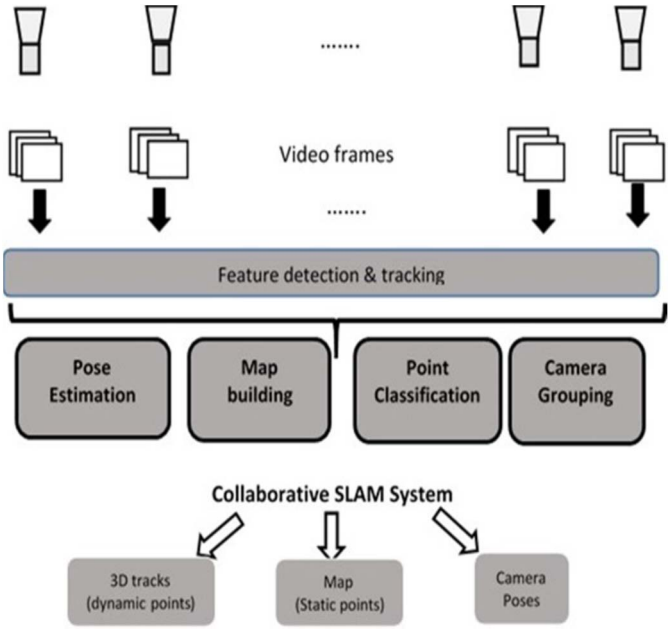


Fig 6. A Proposed COSLAM system architecture [13]

The experimental results showed that the Collaborative Visual Simultaneous Localization and Mapping (CoVSLAM) system was capable effectively calculating position in extremely dynamic and static environment. Collaborative Visual Simultaneous Localization and Mapping (CoVSLAM) was successful in handling with lighting and noise condition in an indoor location. However Collaborative Visual Simultaneous Localization and Mapping (CoVSLAM) was using twelve camera sensor map and localize environment, the performance of the system dropped because of heavy computation of sensors

III. OPEN ISSUES, CHALLENGES AND FUTURE RESEARCH IN VSLAM

Visual Simultaneous Localization and Mapping (VSLAM) over the years have receive a lot of attention and tremendous achievement has been attained. But there are still many issues limiting the full acceptance of this technology. Take for instance the issue of dynamic environment complain by many researchers has not been fully resolved [9], [13]. Dynamic features are object capable of changing their location as a result of external force acting on it. Thus, success attained in this situation is associated with how dynamic the environment is. The more dynamic the environment, the more challenging it becomes to address [9]. Another issue that has not been fully addressed is the illumination variance which mostly affect the image characteristic, with an effect that violate assumption making localization impossible. But in the static environment the effect of Illumination variance has been successful to some extent but still lacks accuracy [7]. However, it is more challenging to address in dynamic environment [9]. Furthermore, some researchers reviewed in this study complain about regular patterns [8], [3]. The effect of regular patterns on Visual Simultaneous Localization and mapping (VSLAM) algorithm can lead to mismatches and missing features which will result in obstacle either been absent or dislocated within the map created for the environment [8]. Lastly, the issue of computation complexity has related speed is often a problem complain by some researchers [13], [3], although researchers tend to reduce the Visual Simultaneous Localization and mapping (VSLAM) computational cost but result attained is successful to some extent [3]. Nevertheless, more research still needs to be done in this area. In future, SLAM technique must be able to eliminate all current challenges to attain maximum effectiveness and efficiency.

IV. CONCLUSION

Visual Simultaneous Localization and Mapping (VSLAM) has been a common method among robotic applications because of its ability to enable an autonomous robot to construct a model of its environment and estimate its position at the same time with the aid of vision sensor such as camera etc. This is an important functionality necessary for a robot to navigate in an unfamiliar area. Numerous researchers have researched in an area of Visual Simultaneous Localization and mapping (VSLAM) with remarkable results. However the review conducted in this study has discovered that challenges such as variations in illumination, specular reflection in environment, inconsistent point clouds, un-textured objects like white walls, regular patterns, reflective surfaces, direct sunlight or shadows illumination, loop closure detection and multiple camera sensors on the robot has the potential to limit or cause the Visual Simultaneous Localization and Mapping (VSLAM) technique to fail whether in a static or dynamic environment. In our effort to improve the Visual Simultaneous Localization and Mapping (VSLAM) technique, we plan to investigate the challenge of environmental noise such as shadow or light intensity/ illumination in our future work because this is often common as a result of high intensity of sunlight in our continent.

REFERENCES

[1] Ramos, F. T., Nieto, J. & Durrant-Whyte, H. F. (2007). Recognising and modelling landmarks to close loops in outdoor slam. In *Proceedings 2007 IEEE International Conference on Robotics and Automation.* p. 2036-2041. IEEE

[2] Clipp, B., Zach, C., Lim, J., Frahm, J. M. & Pollefeys, M. (2009). Adaptive, real-time visual simultaneous localization and mapping. In *Applications of Computer Vision (WACV), 2009 Workshop on.* p. 1-8. IEEE.

[3] Gao, X., & Zhang, T. (2015). Robust RGB-D simultaneous localization and mapping using planar point features. *Robotics and Autonomous Systems*, *72*, 1-14.

[4] Moreno, F. A., Blanco, J. L., & Gonzalez, J. (2009). Stereo vision specific models for particle filter-based SLAM. *Robotics and Autonomous Systems*, *57*(9), 955-970.

[5] Milford, M. J. & Wyeth, G. F. (2008). Single camera vision only SLAM on a suburban road network. In *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on.* p. 3684-3689. IEEE.

[6] Park, C. & Song, J. B. (2015). Illumination change compensation and extraction of corner feature orientation for upward-looking camera based SLAM. In *Ubiquitous Robots and Ambient Intelligence (URAI), 2015 12th International Conference on.* p. 224-227. IEEE.

[7] Al-Mutib, Khalid N., Et Al. "Stereo vision SLAM based indoor autonomous mobile robot navigation." *Robotics and Biomimetics (ROBIO), 2014 IEEE International Conference on*. IEEE, 2014.

[8] Brand, C., Schuster, M. J., Hirschmüller, H., & Suppa, M. (2014, September). Stereo-vision based obstacle mapping for indoor/outdoor SLAM. In *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on* (pp. 1846-1853). IEEE.

[9] Wang, Y., Huang, S., Xiong, R., & Wu, J. (2016). A framework for multi-session RGBD SLAM in low dynamic workspace environment. *CAAI Transactions on Intelligence Technology*, *1*(1), 90-103.

[10] Raguram, R., Chum, O., Pollefeys, M., Matas, J., & Frahm, J. M. (2013). USAC: a universal framework for random sample consensus. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 2022-2038.

[11] Lin, L. H., Lawrence, P. D., & Hall, R. (2013). Robust outdoor stereo vision SLAM for heavy machine rotation sensing. *Machine vision and applications*, 1-22.

[12] Wan, E. A., & Van Der Merwe, R. (2000). The unscented Kalman filter for nonlinear estimation. In *Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000* (pp. 153-158). Ieee.

[13] Zou, D., & Tan, P. (2013). Coslam: Collaborative visual slam in dynamic environments. *IEEE transactions on pattern analysis and machine intelligence*, *35*(2), 354-366.

[14] Thrun, S. (2002, August). Particle filters in robotics. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence* (pp. 511-518). Morgan Kaufmann Publishers Inc.

[15] Thrun, S., & Montemerlo, M. (2006). The graph SLAM algorithm with applications to large-scale mapping of urban structures. *The International Journal of Robotics Research*, *25*(5-6), 403-429.

Virtual Private Network's Impact on Network Performance

Claude Mukatshung Nawej
Dept. of Computer Systems Engineering
Tshwane University of Technology
Pretoria, South Africa
nawejmc@tut.ac.za

Shengzhi Du
Dept. of Mechanical Engineering and Mechatronics
Tshwane University of Technology
Pretoria, South Africa
dus@tut.ac.za

***Abstract*— We categorized VPN's impact in to two different aspects; The affecting aspect, and the affected aspect. The affecting aspect of the impact encompasses factors such as security, algorithms, hardware, and software. Whereas the affected aspect is of the likes of network performance. While VPNs have managed to integrate security, one of the affecting aspect of the impact on one hand, on the other hand, VPNs should be regarded as a potential threat to network performance. In this study, for affordability purpose, we choose to use NS-2 simulated test-bed to shed light on the VPN's performance impact in a network. Considering the most common network performance metrics, throughput and delay; we assessed these performance metrics by means of average and percentage changes theories. The findings emphasize quantitative impact on the TCP/IP throughput than on its counterpart UDP/IP. We finally developed an analytical equation to model this VPN's performance impact.**

Keywords*— *Network performance metrics, NS-2, simulation, Polynomial model, VPN.

I. INTRODUCTION

Almost everyone connects to a network, and the internet is the biggest computer network that exists in this world. The internet encompasses many networks. However, how are networks connecting to each other? One of the solutions to this problem is a virtual private network (VPN). A VPN is a network device or series of devices used to interconnect various networks together [1]. Beside connectivity, VPNs also offer benefits such as flexibility, and security at a cheap cost. On one hand, VPNs offer numerous benefits like flexibility, connectivity, and security at the cheaper price. On the other hand, the computation that provides these adequate benefits however affects the network performance. This weakness prompted us to conduct a study on the impact of the VPN on the network performance. Figure 1 gives us an idea of two networks connected by VPNs.

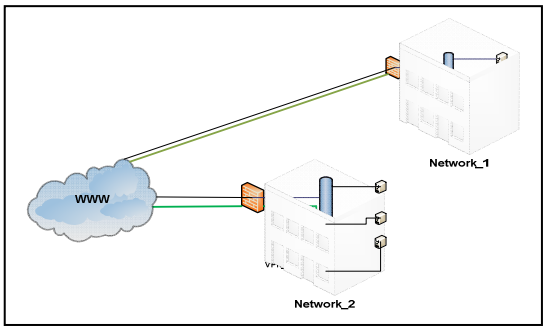


Fig. 1. Two computer's networks interconnected by VPNs

VPN emerges as a key to cost-effective and secure remote access over the internet. The host-to-host data communication over the internet is achieved by the transport layer compatibility [2]. This layer is using two major protocols to move data from source to destination, Transmission Control Protocol (TCP/IP), and the User Datagram Protocol (UDP/IP). TCP uses a congestion window to check and determine how many packets it can send or has sent at one time; UDP on the contrary has no mechanism in place to check loss or delivery; it assume all sent packets are received regardless of the route used.

The remaining of the paper is organized as follows: Section 2 briefly reviews the related work with focus on VPN's impact in general. We underline the protocol configuration parameters of the simulation in Section 3. Section 4 provides details of the simulation test-bed, which is built up in using the NS-2 simulator. Section 5 speaks about collection, analysis of data, and modelling of the impact. Finally section 6 concludes the research, and gives direction for future work.

II. RELATED WORKS

Network performance metrics have two properties namely qualitative and qualitative. Qualitative is a property such as projection of metric, estimation, QoS, behaviour; and quantitative here is a property such as throughput, delay, and lost. VPN's performance impact studies have involved both qualitative and quantitative network performance metrics [3] [4] [5]. In [6], the author underlined in his study that the use of VPN increased the security and at the same time, it decreased the network performance. In [7], the use of a VPN for database transactions affected the delay performance of the query by 446%. On another study using OPNET, the author underscored the impact of a VPN on the response time and load [8]. Performance impact was investigated in a project using an IPsec based VPN with TCP and UDP protocols in [9]. The author found that the use of this VPN affected the packet loss and the network throughput. In [10], the study demonstrated that even an unsecured VPN affected the network performance. Most studies exploited the qualitative properties and only few researches tackled performance metrics quantitatively.

Therefore, in this research, where there is mostly VPN's theoretical discussions and few numbers based on the common performance metrics, we attempt to address quantitatively the question of whether or not the integration of a VPN in a network will affect common network performance. Through NS-2 simulation, we used the major transport protocols together with the common network performance metrics as the variants of the research.

III. PROTOCOL VARIABLES

The transport protocol supports many configuration variables. For many simulations, few of the configuration parameters require modification. Changes to these parameters can have a profound effect on the behaviour of the protocol. TCP/IP and UDP/IP are the most used protocols, and therefore used in this research as follows:

A. TCP/IP

The more commonly modified parameters for TCP/IP protocols are window size and packet Size. For application sitting on TCP, we used the FTP(File Transfer Protocol). Table 1 shows a variation in combination of TCP parameters as used in the simulation.

TABLE I. TCP PARAMETERS USED FOR THE SIMULATION

Window size (Kbytes)	**Packet size (Bytes)**					
	64	128	256	512	1024	2048
10	√	√	√	√	√	√
50	√	√	√	√	√	√
100	√	√	√	√	√	√

B. UDP/IP

UDP has gained popularity recently [11]. For this research, we use CBR; one of the NS-2 applications that sits on UDP protocol. UDP is mainly concerned with packet size expressed in bytes and transfer rate expressed in Mb [12]. In Table 2, we presented UDP combined parameters as used in our test bed simulation.

TABLE II. CBR PARAMETERS USED FOR THE SIMULATION

CBR transfer rate (Mb)	**Packet size (Bytes)**		
	512	1024	2048
10	√	√	√
100	√	√	√
1000	√	√	√

IV. SYSTEM ARCHITECTURE

With the aim to set up a simulation test bed, NS-2 version 3.5 simulation is used as the main tool for the generation and collection of data in this research [13], [14]. Two scenarios A and B were involved.

A. Scenario without VPN (scenario A)

In the scenario A, communication between the two nodes is carried out without VPN as per Figure 2.

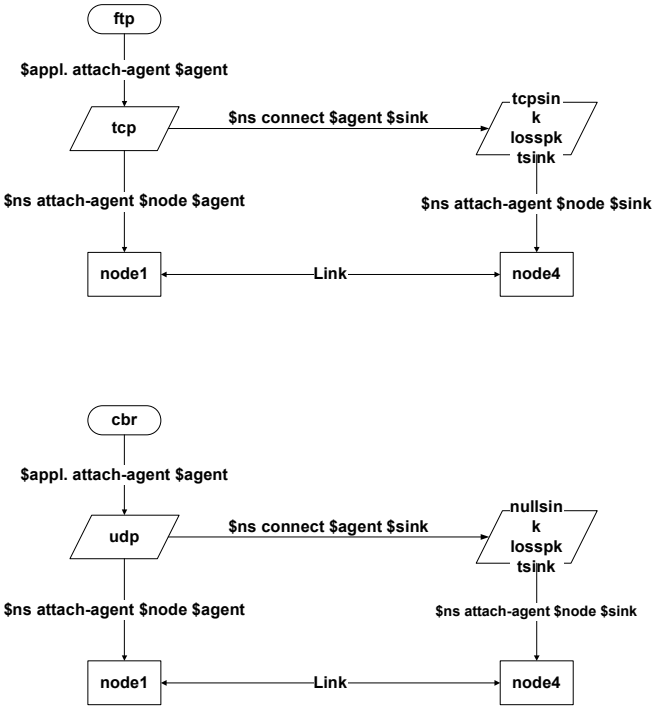


Fig. 2. Scenario without VPN.

B. Scenario with VPN (scenario B)

In the second scenario, communication between the two nodes is carried out with two routers (VPN) in between the initial two nodes. Scenario A differs from scenario B by the use of two extra nodes with routing activated which serve as routers or VPN. See Figure 3.

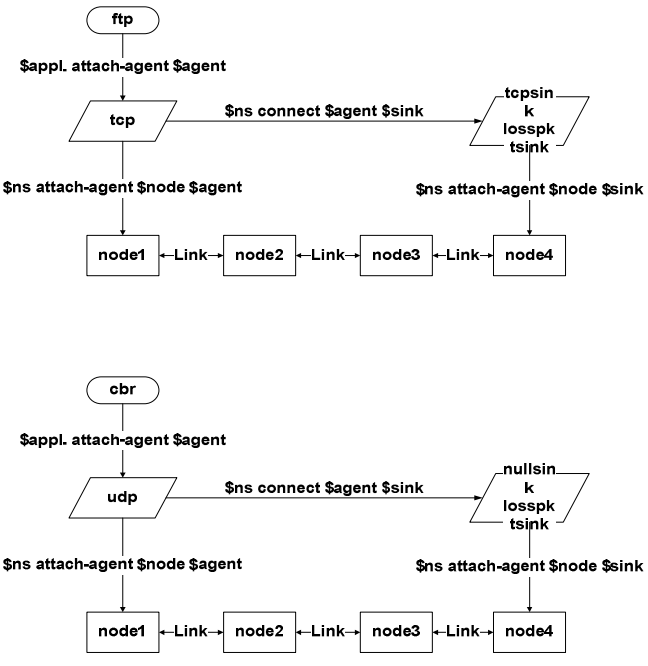


Fig. 3. Scenario with VPN

V. ANALYSIS AND RESULTS

Reference [15] states that the most common network performance metrics are throughput, delay, and loss. Throughput (expressed in Bytes/second) can be defined as a measure of data transfer per unit time between a sender and a receiver. Delay (expressed in millisecond), is the time needed for a packet to travel. In addition, packet loss may be measured as the percentage of frames that should have been forwarded by a network but were not.

In this study, considering the most network performance metrics and taking throughput as being a function of loss, we measure the performance impact with regard to throughput and delay only. To evaluate the impact of VPN on the network performances, a simulation study involving the two major transport protocols, TCP/IP and UDP/IP is launched. We ran the two different scenarios, A and B as defined above, collected data, and average values [16] [17] calculated. Following are the averaged data per protocol.

A. TCP/IP

In this section, under different combination of window size versus packet size, we present using formulae (1) below, statistics of the calculated TCP average throughput and delay.

$$f(av) = \frac{1}{n}\sum_{i=1}^{n}(x_i) \quad (1)$$

Where:

$f(av)$ = average (or arithmetic mean)

n = the number of terms

x_i = the value of each individual item in the list of numbers being averaged.

Further, we discuss average TCP throughput and delay data collected from the experimental testbed.

1) Average TCP throughput: Following are the averaged TCP throughput values calculated from the data simulation testbed collection.

TABLE III. AVERAGED CALCULATED TCP THROUGHPUT DATA

Packet size	Window size					
	10 Kbytes		50 Kbytes		100 Kbytes	
	A	B	A	B	A	B
64 Bytes	0,34	0,11	1,70	0,55	3,39	1,09
128 Bytes	0,55	0,18	2,75	0,89	5,47	1,75
256 Bytes	0,98	0,32	4,84	1,56	9,64	3,09
512 Bytes	1,82	0,60	9,03	2,92	17,98	5,76
1024 Bytes	3,50	1,14	17,32	5,55	34,51	10,96
2048 Bytes	6,83	2,24	33,86	10,89	67,45	21,51

To analyze the results, formulae (2) below is used to work out the percentage changes from information provided in table 3.

$$f(x) = \left(\frac{B-A}{A}\right) 100\% \quad (2)$$

Where:

$f(x)$ = percentage change

For $f(x)$ value positive, this implies positive change (increase), otherwise negative change (decrease)

Due to space limitation, we randomly present graphs of only packet size 512, 1024 and 2048 bytes

a) Window size 10 Kb: : Figure 4 shows the averaged TCP throughput graph of window size 10 Kb using packet size of 512, 1024 and 2048 Bytes.

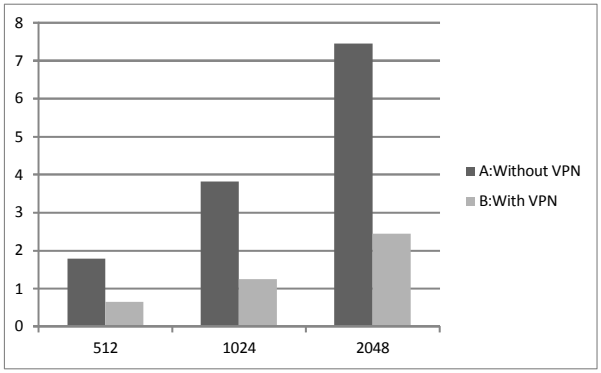
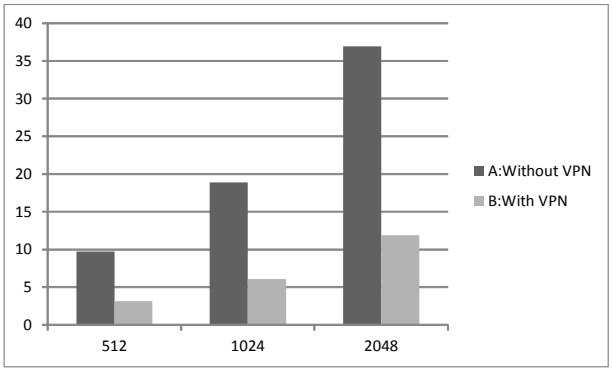


Fig. 4. Averaged TCP throughput of window size 10 Kb.

From figure 4, it is clear that the average throughput of scenario A is bigger than scenario B in all the three packet size cases, $f(x) < 0$. Therefore, decrease in throughput performance.

b) Window size 50 Kb: From figure 5, all the three packet size cases show that scenario A is bigger than



scenario B average wise.

Fig. 5. Averaged TCP throughput of window size 50 Kb.

c) Window size 100 Kb: The average TCP throughput graph of window size 100 Kb, packet size 512, 1024 and 2048 Bytes is plotted in figure 6.

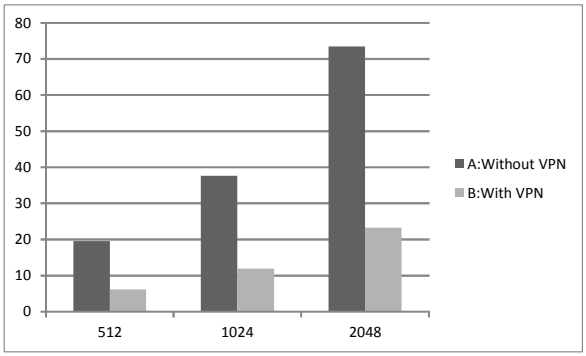


Fig. 6. Averaged TCP throughput of window size 100 Kb..

In all the cases, it is noticed that f(x) < 0. The throughput performance is affected negatively. Thus, decrease in throughput. Besides, it is noticed that throughput performance is function of packet size and window size respectively.

2) Averaged TCP delay: This section introduces the TCP delay of the experiment. Table 4 below provides the averaged TCP delay values derived from the collected simulation testbed data.

TABLE IV. AVERAGED CALCULATED TCP DELAY DATA

TCP Delay (window size: 10 Kb)					
TCP delay Packet size: 512		TCP delay Packet size: 1024		TCP delay Packet size: 2048	
Without Vpn	WithVpn	Without Vpn	WithVpn	Without Vpn	WithVpn
0,010024002	0,0298527	0,010044783	0,029912	0,010086346	0,030035
TCP Delay (window size: 50 Kb)					
TCP delay Packet size: 512		TCP delay Packet size: 1024		TCP delay Packet size: 2048	
Without Vpn	WithVpn	Without Vpn	WithVpn	Without Vpn	WithVpn
0,010025329	0,0300306	0,010047352	0,030096	0,01009141	0,030228
TCP Delay (window size: 50 Kb)					
TCP delay Packet size: 512		TCP delay Packet size: 1024		TCP delay Packet size: 2048	
Without Vpn	WithVpn	Without Vpn	WithVpn	Without Vpn	WithVpn
0,010026996	0,0300583	0,010050577	0,030129	0,010097765	0,030271

From the results in the table above, we analyze same wise as we did with the throughput in the previous section. For not to be repetitive, we discuss just one case, case of window size 10 Kb, from which we generalize results from.

Figure 7 shows the averaged TCP delay performance graph of window size 10 Kb using packet size of 512, 1024 and 2048 Bytes.

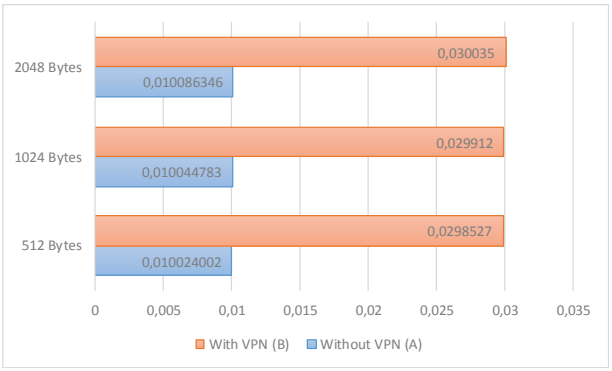


Fig. 7. Averaged TCP delay of window size 10 Kb..

It is noticed that $f(x) > 0$. The change in delay performance is positive. Therefore, increase in delay. Besides, the three window sizes present patterns, which are proportional to their packet sizes respectively. Thus, performance delay function of packet size and window size

B. UDP/IP

In this section, with various combination of packet sizes against transfer rates, we present and discuss statistics of UDP throughput and delay performance. The packet sending rate changes from 10 Mb, 100 Mb and then 1000 Mb. Packet size passes from 512 bytes, 1024 bytes and then 2048 bytes. Throughputs and delays are computed and averages calculated.

1) Averaged UDP throughput: Table 5 shows the averaged CBR throughput of the three different packet sizes: 512, 1024 and 2048 Bytes.

TABLE V. AVERAGED CALCULATED UDP THROUGHPUT DATA

Transfer rate (Mb)	Packet size					
	512 Bytes		1024 Bytes		2048 Bytes	
	A	B	A	B	A	B
10	0,00730	0,00730	0,01450	0,01450	0,02909	0,02909
100	0,00730	0,00730	0,01450	0,01450	0,02909	0,02909
1000	0,00730	0,00730	0,01450	0,01450	0,02909	0,02909

Although the throughput is function of packet size, scenarios A are equal to scenarios B. No change is noticed.

Therefore, no impact on UDP throughput. The transfer rate does not have any influence on the UDP throughput.

2) Averaged UDP delay: This section addresses the UDP delay of the experiment. Table 6 provides the UDP delay values calculated from the data simulation collected from our testbed.

TABLE VI. AVERAGED CALCULATED UDP DELAY DATA

Transfer rate (Mb)	Packet size					
	512 Bytes		1024 Bytes		2048 Bytes	
	A	B	A	B	A	B
10	0,01004	0,01671	0,01008	0,02012	0,01013	0,02218
	A	B	A	B	A	B
100	0,01004	0,01671	0,01008	0,02012	0,01013	0,02218
	A	B	A	B	A	B
1000	0,01004	0,01671	0,01008	0,02012	0,01013	0,02218

As seen from Table 6, UDP delay observations provide the same insights for all its transfer rates per packet size. Therefore, we select a random transfer rate of 1000 MB and present its results graphically in Figure 8 below.

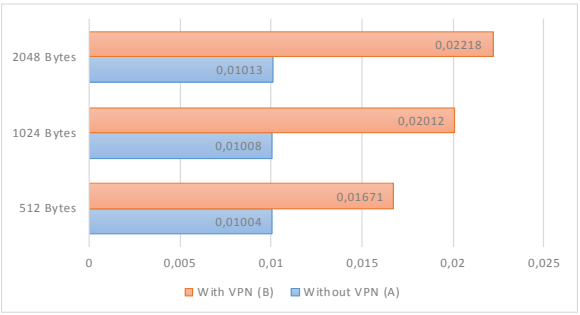


Fig. 8. Averaged UDP delay of the three packet sizes.

$f(x) > 0$, thus increases in delay. Therefore, positive impact

C. Modelling of VPN impact on TCP throughput

The below figures show various instant experimental TCP throughput simulation data. We do the modelling based on the randomly selected packet size of 512, 1024, and 2048.

TABLE VII. INSTANT TCP THROUGHPUT OF WINDOW SIZE 10 KB

TCP throughput			TCP throughput		TCP throughput	
Packet size: 512			Packet size 1024		Packet size 2048	
Time	A	B	A	B	A	B
8	0	0	0	0	0	0
8,5	1,979	0,48	3,81	0,92	7,48	1,8
9	2,208	0,79	4,26	1,53	8,35	3,01
9,5	2,208	0,71	4,26	1,36	8,02	2,67
10	2,208	0,71	4,26	1,36	8,35	2,67
10,5	2,208	0,79	4,09	1,36	8,35	2,67
11	2,208	0,71	4,26	1,53	8,35	3,01
11,5	2,208	0,71	4,26	1,36	8,15	2,67
12	2,208	0,79	4,26	1,36	8,22	2,67
12,5	2,12	0,71	4,26	1,53	8,35	2,67
13	0,088	0,71	4,26	1,36	8,35	3,01
13	0,088	0,71	4,26	1,36	8,35	3,01

TABLE VIII. INSTANT TCP THROUGHPUT OF WINDOW SIZE 50 KB

TCP throughput			TCP throughput		TCP throughput	
Packet size: 512			Packet size 1024		Packet size 2048	
Time	A	B	A	B	A	B
8	0	0	0	0	0	0
8,5	8,94	1,43	17,23	2,76	33,34	5,41
9	11,04	3,97	21,28	7,66	40,89	13,73
9,5	11,04	3,53	21,13	6,81	41,43	14,67
10	11,04	3,53	20,84	6,81	41,76	13,36
10,5	11,04	3,97	21,02	6,81	41,76	13,36
11	11,04	3,53	21,28	7,66	41,06	14,37
11,5	10,95	3,53	21,28	6,81	40,92	14,03
12	10,8	3,65	21,28	6,81	41,63	13,36
12,5	10,93	3,85	21,28	7,66	41,76	13,36
13	10,04	3,53	21,28	6,81	41,69	15

TABLE IX. INSTANT TCP THROUGHPUT OF WINDOW SIZE 100 KB

TCP throughput			TCP throughput		TCP throughput	
Packet size: 512			Packet size 1024		Packet size 2048	
Time	A	B	A	B	A	B
8	0	0	0	0	0	0
8,5	17,01	1,997	32,67	3,85	62,21	7,55
9	22,08	7,95	42,12	14,78	82,65	27,09
9,5	22,08	7,07	42,12	14,16	83,19	29,7
10	22,08	7,07	42,12	13,62	82,85	26,73
10,5	22,03	7,9	42,3	13,62	82,68	26,73
11	21,84	7,11	42,56	15,32	82,65	27,73
11,5	21,84	7,07	42,56	13,62	82,68	29,06
12	21,84	7,19	42,56	13,62	83,35	26,73
12,5	21,97	7,83	42,56	15,32	82,68	26,73
13	22,08	7,07	42,49	13,62	82,65	28,36

From the above, a relationship between the two scenarios is developed. A square coefficient of the polynomial model is found to be higher than it was in the other models. For not to be exhaustive, we selected a random 1024 Kb packet size of instant TCP throughput 50 Kb window size from Table VIII to develop the relationship. A polynomial model fits best our experimental data. From this, a prediction graph is provided in Figure 9. The Y-axis represents the throughput with VPN in between, and X-axis represents the throughput without VPN in use.

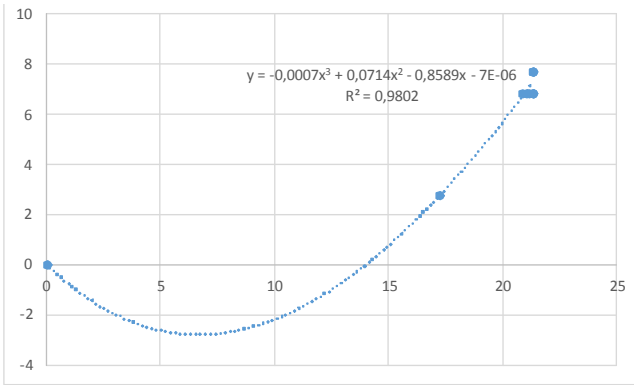


Fig. 9. Predicted values in a polynomial model of order 3.

From Figure 9, we identify equation (3) with R^2 = 0, 9802

$$y = -0{,}0007x^3 + 0{,}0714x^2 - 0{,}8589x - 7E\text{-}06 \quad (3)$$

Knowing the TCP throughput performance of scenario without VPN, the polynomial model equation in (3) provides a convenient tool to predict the TCP throughput performance impact due to the use of VPN.

VI. Conclusion and Future Work

Our test-bed simulation is considered an ideal environment, where no background and outside parameters to the communication channel get involved. The simulation results under the above circumstances demonstrated that the use of VPNs draws impact on the performance depending on the type of protocol in use. When VPN was used on the network simulation with TCP protocol, throughput decreased and delay increased. The patterns showed that the VPN's impact on both the TCP's throughput and delay performance is function of packet size. We could reduce the impact in favouring the use of small packet size, and window size respectively. For UDP throughput, it was proven that there is nothing lost in the journey; Investigation revealed that UDP throughput was unaffected by the use of VPN. All traffic injected was received regardless of the packet size and transfer rate. For the delay, it was discovered instead that the UDP delay was sensitive to the VPN and this UDP delay was function of only packet size, and not function of transfer rate. For future work, we will investigate VPN impact on network performance in a non-ideal environment, where the sender and receiver sharing the same channel communication, and where the two protocols are running concurrently each next to the other. In addition, we will try to answer to the questions: what are the dangers of decreasing the TCP window size, and packet size; how much is too much?

References

[1] N.K. Tan, *Building VPNs with IPsec and MPLS*. New York, USA: McGraw-Hill networking, 2003.

[2] A. Sikora, D. Lill, and V. Z. Groza, "Comparison of internet and dedicated communication protocols for instrumentation and measurement systems using short range wireless networks," in *Proceedings of IEEE Instrumentation and Measurement Technology Conference*, 2006, pp. 982-985

[3] A. Passito, E. Mota, R. Aguiar, L. S. Carvalho, E. Moura, E. Nascimento, and I. Biris, "Evaluating voice speech quality in 802.11b networks with VPN / IPsec," in *Proceedings of IEEE 13th International Conference on Networks jointly held with 7th Malaysia International Conference on Communications*, 2005, pp. 161-165

[4] A. K. Agarwa and W. Wang, "Measuring performance impact of security protocols in wireless local area networks," in *Proceedings of IEEE 2nd International Conference on Broadband Networks*, 2005, pp. 581-590.

[5] R. Weerawarna, "TCP/UDP network performance evaluation of various IPsec algorithms,"Masters dissertation, Dept. Comp., Unitec institute of technology, New Zealand, 2013.

[6] A. Malik, H. Verma, and R. Pal, "Impact of firewall and VPN for securing WLAN," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol.2, no.5, pp. 407-410, 2012.

[7] J. M. Powell, "The impact of virtual private network (VPN) on a company's network,"Honors dissertation, Dept. Man. Inf. Syst., Utah state university, Logan, UT, 2010.

[8] H. Bourdoucen, A. Al Naamany and A. Al Kalbani, "Impact of implementing VPN to secure wireless lan," *International Journal of Computer, Information and Systems Science*, vol.3, no.1, pp. 17, 2009.

[9] S. Pruthi, and G. Pruthi, "Performance evaluation of wireless IPsec VPN," *International Journal of Information Technology and Knowledge Management*, vol.2, no.1, pp. 41-44, 2009

[10] M.T. Ashraf, J.N. Davies, and V. Grout: "An investigation into the effect of security on performance in a VoIP network," in *Proceedings of 5th Collaborative Research Symposium on Security, E-Learning, Internet and Networking*, 2009, pp.15-28.

[11] M. Zhang, M. Dusi, W. John, and C. Chen: "Analysis of UDP traffic usage on internet backbone links," in *Proceedings of IEEE 9th Annual International Symposium on Applications and the Internet*, 2009, pp. 20-24.

[12] L. Finn, R. Mulholland, and G.J. Gibbon: "Design and implementation of a prototype computer based rehabilitation device for the lower extremities," *SAIEE Africa Research Journal*, vol.95, no.1, pp.29-32, 2003.

[13] T. Issariyakul and E. Hossain, *Introduction to network simulator ns2*, 2nd ed. New York, USA: Springer, 2008.

[14] ns-2. (2014). *The Network simulator* [Online]. Available: http://www.isi.edu/nsnam/ns

[15] M. Hasib, "Analysis of packet loss probing in packet networks," Ph.D. dissertation, Dept. Elect. Eng., Queen mary university of London, London, UK, 2006.

[16] G. Keller, *Statistics for Management and Economics*, 10th ed. Australia: Cengage Learning, 2014.

[17] S. M. Ross, *Introductory statistics*, 3rd ed. Canada: British Library, 2010.

Proposing a Blockchain-based Solution to Verify the Integrity of Hardcopy Documents

Sthembile Mthethwa, Nelisiwe Dlamini, Dr. Graham Barbour
Council for Scientific and Industrial Research (CSIR)
Modelling and Digital Science Unit, Information Security
Pretoria, South Africa
{smthethwa, ndlamini2, gbarbour}@csir.co.za

Abstract—**Even with the ability to produce documents digitally, the paperless environment has yet to become a reality in South Africa. Hardcopy documents are still printed daily which makes them susceptible to document fraud. In South Africa, a case was reported recently, where someone who was creating fake documents was exposed. This introduces the challenge when using hardcopy documents which is loss of integrity. Thus, it is vital to have systems in place to verify document integrity and be able to determine when a document has been tampered with. Various techniques have been used to secure documents, yet the challenge persists. The combination of 2D barcodes, digital signatures, Optical Character Recognition (OCR), cryptographic hashing has proved the potential to achieve good results when combined. Recently, blockchain has been added as one of the techniques to be employed for document verification. This paper presents a proposed solution that incorporates the combination of 2D barcodes, OCR, cryptographic hashing and blockchain. As this is still on-going work, experiments are still required to demonstrate the viability of the solution.**

Keywords—blockchain, 2D barcodes, cryptographic hashing, digital signatures, document generation and validation, document verification, integrity, optical character recognition (OCR), secure hash algorithm (SHA-256), tesseract

I. INTRODUCTION

Nowadays digital documents have become a large part of every sector whether public or private resulting from the transformation of modern technology. This not only allows the dissemination of information but also preserves these documents in digital form and promotes paperless environments. However, a variety of these documents are printed every day such as; academic certificates, wills, case files, birth and marriage certificates, national identity documents, insurance documents, passports and drivers' licenses, etc. and issued to different people because printed documents are the most prevalent form of trusted communication. A great challenge is therefore experienced along with the advancement in technology; which now make it possible to easily reproduce falsified documents as they are now susceptible to unauthorized alterations [1, 2].

A series of forgery cases have been reported in the past. In India, a woman who was applying for a passport presented a falsified birth certificate, and it took some time for the concerned agency to detect the forgery. The police reported this as a common occurrence when it comes to applicants [5]. Incidents of falsified academic documents have also been reported, for instance, in Singapore, where foreign nationals were convicted after being charged with forgery of academic documents [4]. Additionally, in South Africa, two people presented fraudulent asylum documents at their workplace which they claimed were received from a state's agency official [3]. Recently a man, who produced unauthorized pay slips and bank statement and sold them to people, was exposed [6]. The counterfeiting of these documents enabled people to get credit and the credit companies are unable to trace these people after that, which has cost them a lot of money [6]. Such cases indicate the seriousness of hardcopy document fraud and demand a need to augment the solutions that solve this problem by introducing numerous methods to secure the original document from being forged in any way.

Luckily, this downside in the security of hardcopy documents has attracted attention from many researchers, and research in this area is advancing, with the aim to alleviate unauthorized altering and compromised integrity of these documents and the use of falsified documents. This entails proving that a scanned/physical copy of a document is the same as its original document [1, 2]. In doing this, various methods are implemented to record information pertaining to the original document, called the original template; and encode this information in a barcode then insert the barcode into the document and print it. Upon presentation of a copy of the original document, recognition techniques attempt to extract this information from the copy, producing a copy template. The two templates are then compared. Two key issues arise; first is the problem of storing the original template content, and secondly, the problem of extracting the copy template.

While the original template can be stored in a database, an ideal solution would be to store the original template as visible information on the original document itself. To supplement this, there are methods that are used to add information to the original document to ensure that a copy is not tampered with in any way. Amongst these methods are watermarks, document signatures, barcodes, hashing, etc. However, information stored using these methods is limited, for instance all document content cannot be included in a barcode because of the space limitations. Rather than storing the original template on the document, a hash is stored instead [2], and the problem of extracting the stored template from the copy emerges as a hash maps the document content to a string that represents the content, it does not contain the actual content in the original document.

Despite this, the use of hash values in blockchain-based methods, is undoubtedly a reliable solution that is now applied in document verification systems [7]. The blockchain which is

a distributed, replicated and synchronized public ledger has made it possible to implement solutions that validate the integrity of the documents issued. However; more work still remains, the transition from eliminating the use of hardcopy documents to using digital documents hasn't been successfully navigated as yet considering that hardcopy documents are still widely used. In this paper we use the standard Optical Character Recognition (OCR) technique and incorporate the use of blockchain technology to present an agnostic solution that focuses on the ability to verify the integrity of both digital and hardcopy documents.

This paper is organized as follows, in Section 2, we present the literature review. Then a discussion of the proposed solution is presented in Section 3 and Section 4, concludes the study.

II. Literature Review

Over the years, technology has made it very easy to produce digital documents which are easy to retrieve, access and store, and encourages the change to more paperless environments. However; printed documents are still predominant and used to serve the purpose of communicating relevant information to people even though these documents have been perceived as cumbersome and inefficient [8, 9]. With the advancement of technology, the demand to verify important hardcopy documents has escalated, as the issue of fraudulent documents continues to aggravate. Falsifying a hardcopy document requires less effort these days, because these documents are inherently insecure and most have no passwords or digital signatures unlike digital documents. A long list of techniques have been proposed to mitigate the problem of forged documents mostly digital documents. But research in the area of securing hardcopy documents is starting to gain a lot of traction, since the use of these documents has become undeniable [1].

Some of the prevalent techniques include the use of watermarking, which aims to preserve the integrity of a document. Watermarking can either be in a digital or printed format [1]. This technique is still vulnerable to attacks, which may not necessarily remove the watermark imprinted, but rather disable its readability, the success of watermarks also relies on high quality printers, which incurs cost [8]. Nevertheless, it remains an active research area and continues to be improved [10]. The use of OCR, to recognise text from an image file is also prevalent. OCR is the best tool with regards to character recognition, whereby it takes in an image and returns the recognised text. Tesseract is quiet popular as an open source OCR tool, and is identified as having better accuracy and precision than other OCR tools, e.g. Transym OCR and GOCR [11]. The main issue with using OCR independently, is that it is not sufficiently reliable to determine the accuracy and is not generally 100% especially when a document has text that is not solidly black and a noisy white background, nonetheless it can be trained to achieve the expected accuracy [11, 12, 13].

Cryptographic techniques such as cryptographic hashing, Public Key Infrastructure (PKI) and digital signatures, are also very common in document verification, matter of fact it has been used in many studies to secure documents. [10] presented a solution to prove the authenticity of a document and verify it, using digital signatures. They also considered incorporating blockchain technology but decided using PKI digital signatures was sufficient for their system. Blockchain technology is flourishing in this area, its properties such as immutability, transparency and authenticity of digital records; has attracted it to a number of private and public sectors which have welcomed its use to counter document fraud [14]. Civic is one of the companies that have successfully implemented secure identity verification using blockchain technology, for this system to work, cryptographic hashing which plays a major role in Blockchain-based solutions, is employed [14]. Academic institutions have also adopted Blockchain use, e.g. Massachusetts Institute of Technology (MIT) is one of many institutions that now uses the blockchain to register digital educational certificates and allows people to authenticate these certificates, also applying vast use of cryptographic hashing for verification [6, 15]. Another system, Stampery uses a combination of blockchains, to ensure the integrity, existence and the ascription of any file or document, even communication. Once these files have been anchored to the blockchain anyone, anywhere in the world can verify their integrity [16].

Several research studies have also explored the use of two – dimensional (2D) barcodes, whereby information about the document is stored in a barcode and used later for the process of verification [1]. 2D barcodes are commonly used for document verification as they can store more data than 1D barcodes. To strengthen the security of barcodes, various cryptographic techniques are used i.e. PKI, data compression, hash functions, digital signatures [1]. [1] proposed a system whereby, barcodes are used with the help of these cryptographic techniques. Thus, showing the importance of integrating different components to design a suitable solution for the problem of document forgery. A limitation that comes with the use of barcodes is size (the amount of data that could be stored in a barcode) and once a document has numerous barcodes, it starts using a lot of space that can be used for content. Most of the proposed solutions that utilize barcodes store the entire document content in the barcode [17, 18]. In [19] we eliminated this by only storing the information that should be validated in a document, but still encountered the challenge of having numerous barcodes. Hence, in continuing with this ongoing work, the solution proposed in this study aims to decrease the number of barcodes by transferring the information used to verify the hardcopy documents to a blockchain.

From all these studies, and present implementations it can be concluded that efficient techniques must not only be effective but affordable, and implemented well to ensure the security of a document in making sure that unauthorized alterations can be detected. Thus, this study aims to provide an effective, simple and fast method of document integrity verification through the usage of 2D barcodes, OCR, cryptographic techniques and blockchain.

III. Proposed Solution

This section describes the proposed solution for verifying a hardcopy document and a digital document which is an extension of the solution that was presented initially in [19, 20]. In [19] the solution consisted of 4 components, namely; cryptographic hashing, digital signatures, OCR and 2D barcodes. Experiments were conducted using 3 different fonts i.e. Times New Roman, OCRB and AnyOCR and the highest accuracy obtained for AnyOCR was 100% which presented an opportunity to improve the solution so that it can work with different fonts. The documents generated in [20] consisted of 7 barcodes positioned at the bottom of the document, which presents a challenge to those who might want to adopt the system. The number of barcodes is dependent on the information that needs to be stored as each barcode has storage space limitation.

This led to some research on finding ways we can make the solution easily adoptable without the issue of having more barcodes to deal with when documents are generated as this might not necessarily fit in with the company's objective. All of the components used in the previous solution will still be used for this solution except only one barcode that contains information used to verify will be placed on the document. The solution is designed in a way that, if an attacker tries to tamper with the document, the system can detect those changes. In this paper we won't discuss the other components used for this solution as this was done in [20]. Only the added component and the modifications introduced are discussed.

A. Components of the proposed solution

1) Barcodes: This component was used in our previous solution, however; the limitations of storage space led to the use of more barcodes in order to accommodate all the information required to validate a document. The maximum capacity of a version 40 barcode in byte mode is between 1273 – 2953 bytes. In [20] , we observed that this barcode capacity presented flaws as it possessed high resolution which yielded negative results after printing and scanning the document. The quality of the data stored in the barcode was poor. To improve this the metadata was divided into 7 portions and stored in 7 smaller barcodes making sure that the capacity used in the barcodes is distributed equally and doesn't reach a growing rate that presents poor quality when the barcodes are decoded and read. The possibility of the capacity growing in the barcode therefore presents a downfall. This challenge might prevent the adoption of this system as a company would not agree to change their structure in-order to accommodate for more barcodes. For the reason that existing companies already have an acceptable format and layout they use to create documents. We realised when we demonstrated the first developed prototype a lot of questions were centred on the multiple barcodes, and some companies were concerned about the space used by the barcodes and the aesthetics of the document design. This challenge led us to trying other means in order to eliminate the inclusion of numerous barcodes. As a solution, we decided to transfer all the information previously stored in the barcodes to a blockchain and only have one small barcode in a document that would contain less information which will be used to verify the integrity of the document. Having one barcode would make it easier for the solution to blend perfectly with the existing structure of documents without changing it massively.

2) Blockchain: The introduction of blockchain has sparked a lot of interest in the research field. Researchers are constantly looking for means where this technology can be applied. The inception of this technology presented a lot of opportunities in the field and not only is it sparking interest in the field of cryptocurrencies, but its being studied for other purposes as well. Blockchain introduced a decentralized method of storing information, whereby all the participants have a copy of the blockchain. Thus eliminating one single point of failure. There is a plethora of blockchains e.g. Bitcoin, Ethereum, Hyperledger, Ripple, etc., one can choose from depending on what they are trying to achieve. Just like any other technology, blockchain posesses limitations i.e. the size of information one can store in the blockchain. This solution focuses more on the components selection rather than the cost of using each of these components particularly the use of barcodes and blockchain by the companies. The cost of the selected components and affordability of the company will be established accurately during the implementation of the solution, certain measures will be used to determine which blockchain will be used and deployment of this system in different company environments.

B. Proposed Solution Design

The solution consists of 2 main processes; generation and validation process. These processes utilize all the components discussed in the previous solution [20] as well as the ones discussed in the previous sub-section.

1) The Generation Process: This process includes the definition of two types of text; normal and validation text, validation text is hashed to produce a single hash value that is then encrypted with a secret key to obtain a digital signature. This process is illustrated in fig. 1 and 2.

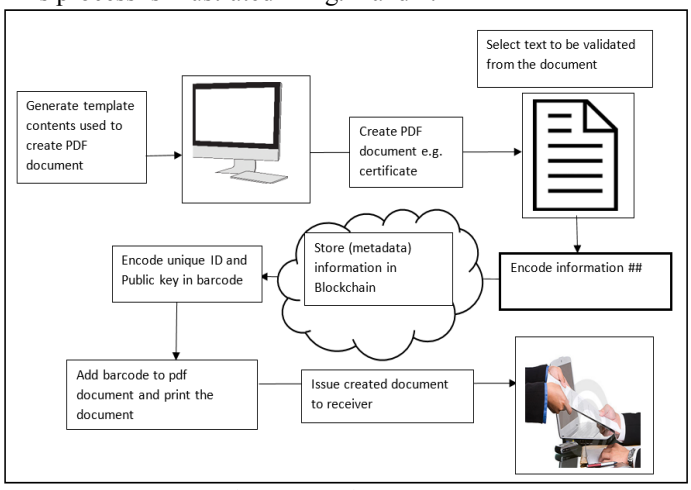


Fig. 1. Generation Process.

The digital signature and metadata are stored on the blockchain and a unique key is generated. This unique key and public key associated with the secret key used to create a digital signature are encoded to the barcode which is placed on a document. The metadata consists of; position, length, width and checksum values that are derived for each validation text, hash produced for all the validation text labels and timestamp of when the document was created. Finally, a digital copy is sent to the recipient and the pdf document is generated, printed and presented.

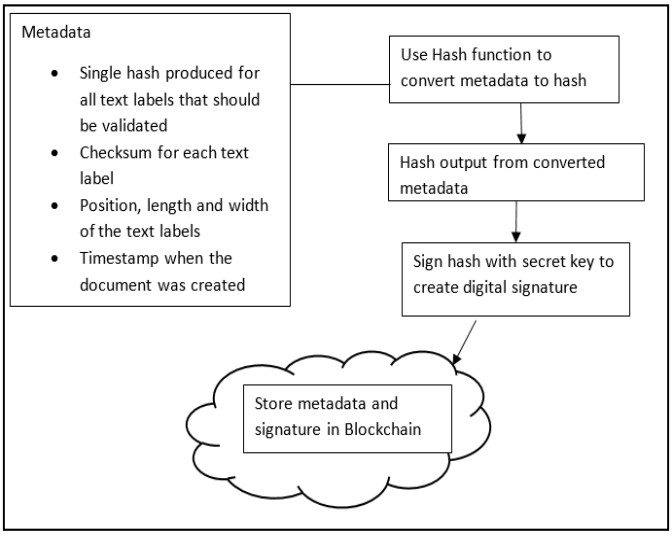


Fig. 2. Encode Information.

2) The Validation Process: Once the documents have been generated, printed and issued, the recipient can then use the documents in various cases, i.e. applying for jobs, applying for bank acounts etc. When these documents are submitted, they must be verified in order to determine whether their integrity has been maintained. Users can either submit a digital copy or a hardcopy document. If a hardcopy document is presented, the document must be scanned first in order to obtain a digital copy. To start the process of validation, the system reads in a scanned image and the barcode is identified and decoded. The barcode consists of a public key and unique key (which is used to fetch metadata related to the document from the blockchain). With the use of the public key extracted from the barcode, the digital signature is validated, if valid the metadata is extracted and used to locate the validated labels in the document. Thereafter, Tesseract OCR is used to validate the text labels. The hash (of all the validated text labels) is calculated and compared with the one from the original document (retrieved from the blockchain). If the comparison fails, it means the document has been altered. In addition to the hash that is included, a checksum for each text label is also calculated, this aids to point the exact text label that is not matching. Fig. 3, illustrates the process of document validation.

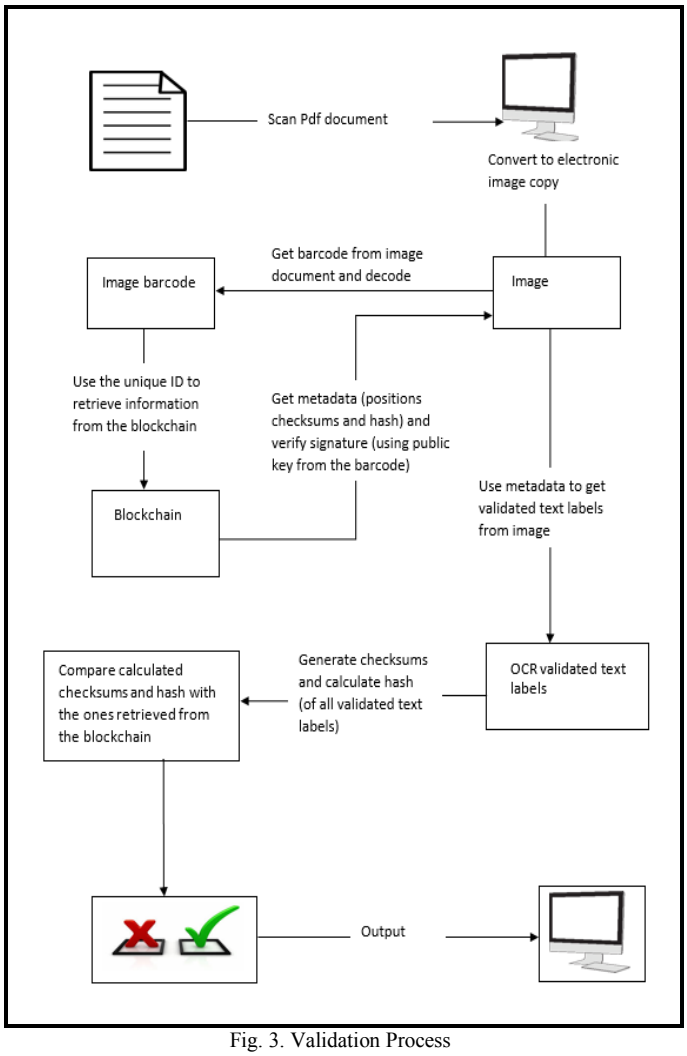


Fig. 3. Validation Process

The evaluation of the proposed solution will use a dataset generated using the generation process. The dataset would consist of 100 generated documents, using different fonts, i.e. Courier, Arial, Times New Roman etc., in order to determine the accuracy of OCR on these documents. The dataset will also be separated into different groups, on the size of text, which are; small, medium and large text. The validation process will be used to verify the integrity of the documents. To evaluate this, the information retrieved from the blockchain must be the same as the one presented when hashed.

Based on the studies conducted, the use of blockchain to verify documents is not something new and the implementations discussed in this paper have been a great success. Be that as it may most proposed and implemented solutions focus more on utilizing a single hash for verification purposes, whereby the document content is hashed and the calculated hash is saved on the blockchain [15]. To augment the existing solutions, our solution uses the blockchain to save more information about the document which is used during the process of validation, and also aims to exclude the irrelevant information in the document. Not only are we intending to identify a document that has been tampered with, but we also want to be able to show where the document has been changed.

This is made possible by including the exact position of the text that should be validated, (x and y coordinates) and the width of the text in the metadata. Before using OCR when the text is extracted from the document all white spaces are removed to minimize any additional white spaces introduced in the process. In most case a single hash value is used to represent the document's content, which cannot be obtained as a hash is designed to be a one way function, whereby we don't know the contents of the document, nor are we aware of the location of the altered text.

IV. Conclusion

This paper presented a proposed solution for the problem of document forgery, which is an extension to our previous proposed solution. The solution employed 4 techniques; OCR, cryptographic hashing, digital signatures and 2D barcodes. OCR was the first technique to be implemented, whereby documents were generated using a font known as AnyOCR (which is designed for OCR tools) and Tesseract was used to validate the documents. The experimental results yielded an accuracy of 100%, which is good. The second part of the experiment was to combine all the techniques, whereby new documents were generated and validation text was specified which was then added to the barcodes that are positioned at the bottom of the documents. Using our validation process, the system was able to detect when documents have been tampered with. This paper extends the previous solution by limiting the number of barcodes used to a single barcode and using blockchain to store the information that was previously stored in barcodes. This proposed solution will be implemented and tested for its practicability to detect forgery and ensure that the integrity and authenticity of a hardcopy document is maintained.

Acknowledgment

The authors would like to express their gratitude to the CSIR Modelling and Digital Science Unit for sponsoring this research.

References

[1] A. Husain, M. Bakhtiari, and A. Zainal, "Printed Document Integrity Verification Using Barcode," *Journal Teknologi (Sciences and Eng*, pp.99-106, 2014.

[2] M.H. Eldefrawy, K. Alghathbar, and M.K. Khan, "Hardcopy document authentication based on public key encryption and 2D barcodes," In *Biometrics and Security Technologies (ISBAST), 2012 International Symposium,* pp. 77-81, IEEE, March 2012.

[3] R. Jain, and D. Doermann, "Visualdiff: Document image verification and change detection," In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on* pp. 40-44, IEEE, August 2013.

[4] N. Ganesan, "*Three foreigners jailed in Singapore for submitting fake academic certificates _ Human Resources Online*, *Human Resources*," Available at: http://www.humanresourcesonline.net/three-foreigners-jailed-for-submitting-fake-academic-certificates/ (Accessed: 07 February 2018), 2017.

[5] S. Kalipa, "*Home Affairs official sold us fake papers _ IOL News*, *Crime and Courts IOL News*," Available at: https://www.iol.co.za/news/crime-courts/home-affairs-official-sold-us-fake-papers-1929048 (Accessed: 08 February 2018), 2015.

[6] C. Lewis, "SABC News exposes fake payslips, bank statements," 8 May 2018, 2018. [Online]. Available: http://www.sabcnews.com/sabcnews/sabc-news-exposes-fake-payslips-bank-statements/. [Accessed: 13-Sep-2018].

[7] B. Cresitello-Dittmar, "Application of the Blockchain For Authentication and Verification of Identity," Independent Paper, 2016.

[8] Y. S. Joshi, "The Future of Enterprise Printing: Securing Hardcopy Documents in the Digital Age: white paper," CIO Insight, no. July, pp. 1–12, 2014.

[9] C. Lakmal, S. Dangalla, C. Herath, C. Wickramarathna, G. Dias, and S. Fernando, "IDStack - The common protocol for document verification built on digital signatures," 2017 Natl. Inf. Technol. Conf. NITC 2017, vol. 2017–Septe, no. September, pp. 96–99, 2018.

[10] S. R. M. Oliveira, M. A. Nascimento, and O. R. Zaiane, "Digital Watermarking: Status, Limitations and Prospects," Technical Report TR 02-01, Department of Computing Science, Alberta University, Edmonton, Alberta, Canada, 2002.

[11] S. Dhiman, and A. Singh, 2013. "Tesseract vs gocr a comparative study," *International Journal of Recent Technology and Engineering*, *2*(4), pp.80, 2013.

[12] C. Patel, A. Patel, and D. Patel, 2012. Optical character recognition by open source OCR tool tesseract: A case study. *International Journal of Computer Applications*, *55*(10), 2012.

[13] V. S. Chandel, "Deep Learning based Text Recognition (OCR) using Tesseract and OpenCV," 06 June 2018, 2018. [Online]. Available: https://www.learnopencv.com/deep-learning-based-text-recognition-ocr-using-tesseract-and-opencv/. [Accessed: 13-Sep-2018].

[14] B. Karanjia, A. G. Karanth, S. Veerapaneni, S. Goswami, A. Sharma, and M. Boda, "Blockchain in the Public Sector – Transforming Government Services through Exponential Technologies," 2017.

[15] Universa, "Blockchain in Education," 23 May, 2018. [Online]. Available: https://medium.com/universablockchain/blockchain-in-education-49ad413b9e12. [Accessed: 04-Sep-2018].

[16] A. S. de P. Crespo and L. I. C. García, "Stampery Blockchain Timestamping Architecture (BTA) - Version 6," 2017, pp. 1–21.

[17] M. Salleh, and T.C. Yew, "Application of 2D Barcode in Hardcopy Document Verification System," In *ISA*, pp. 644-651, June 2009.

[18] C.M. Li, P. Hu, and W.C. Lau, "Authpaper: Protecting paper-based documents and credentials using authenticated 2D barcodes," In *Communications (ICC), 2015 IEEE International Conference,* pp. 7400-7406, IEEE, June 2015.

[19] S. Mthethwa and N. P. Dlamini, "Verifying the Integrity of Hardcopy Document Using OCR," in 2nd International Women in Science Without Borders (WiSWB)-Indaba, 2018.

[20] N. Dlamini, S. Mthethwa, and G. Barbour, "Mitigating the Challenge of Hardcopy Document Forgery," in International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD), 2018, pp. 1–6.

Atmospheric Effects on Free Space Optics Wireless Communication: Applications and Challenges

Solly P Maswikaneng
Dept. Information Technology
Tshwane University of Technology
Pretoria, South Africa
maswikanengps@tut.ac.za

Pius A Owolawi
D*ept. Computer Systems*
Tshwane University of Technology
Pretoria, South Africa
owolawipa@tut.ac.za

Sunday O Ojo
D*ept. Computer Science*
Tshwane University of Technology
Pretoria, South Africa
ojoso@tut.ac.za

Maredi I Mphahlele
Dept. of ICT
Tshwane University of Technology
Pretoria, South Africa
mphahlelemi@tut.ac.za

***Abstract*— Demand for data and multi-media services in recent times has really grown and it is expected to grow more in the next decade with application ranges from high speed real-time/live streaming to live video-conferencing, and so the demand for new and richer data services to meet up with the new and more advanced applications. In other to meet up with these great users' expectations, there is a need to explore other viable access network technologies that can offer services beyond what is already in the offering. Free space optical (FSO) communication system is the predominant choice to replace existing wireless technology with great availability of link capacity and huge bandwidth to meet the user's expectation of the 21st century. This system is reliant on optics signal transmitted in a free space channel, hence termed FSO. This type of communication does not require any licenses for its operation and offers many features being the low-start up and operational cost, rapid development and high fiber-like bandwidth. FSO uses line-of-sight path for the communication between two points and beam of light to provide optical connection that can send and receive video, voice and data information. This technology provides a massive and unregulated bandwidth where a data rate in excess of 100Gbit/s is achievable over a distance of 1-4km. However, it suffers from degradation in the signal quality due to its susceptibility to atmospheric channel impairments which limits its wide deployment and quality of services. This paper explores the FSO system, atmospheric effects of implementing this system and mitigating techniques.**

Keywords—multi-media, real-time, free space optical fiber-like, atmospheric channel

I. INTRODUCTION

The average internet users' need has gone beyond what the traditional internet offers like surfing the web and using email services which is simple in terms of quality as data will be transmitted as quickly as possible using the resources available to them. Currently number of technologies used to meet these users' expectations includes copper-based technologies, hybrid co-axial and optical fibre cables, fibre-to-the-home, broadband RF/microwaves technologies, etc, but these technologies have limitations and so it is becoming clear that they are rather now becoming very inadequate in meeting up with the end-users' demand.

FSO communication is an optical communication technology that uses light (laser beam, visible, infrared and ultraviolet band) propagation in free space or unguided medium via line of sight (LOS) to transmit data wirelessly for the purpose of telecommunication and computer networking [1, 2] and can enable transmission of optical up to 2.5 Gbps of data, voice and video at long distances (up to 4 km) [3]. Free space entails air, outer space, vacuum and similar media. FSO alone is capable of offering a flexible networking solution that provides an easily expandable, affordable, secured, high-speed and free license wireless broadband connectivity for a number of applications such voice, data, video and entertainment [4]. The FSO communication requires no spectrum licensing or is affected by the constraints of the traditional fibre optics system which makes it the better option in meeting the desire for high bandwidth and high capacity applications. This desire has led to a very fast popularization of smart terminals with emerging new applications, future cellular networks i.e. 5G wireless network and beyond [5, 6,7], which is emphasized by a highly communication data rates and large system capacity. The FSO communication ranges from about 100 m to a few kilometres with a throughput rate up to 1.25 Gbps at frequencies above 300 GHz and wavelength ranging typically from 785 to 1550 nm [8].

With the numerous advantages of the FSO communication technology outlined, it has its own challenges, such as geometric loss, atmospheric loss, ambient light source disturbances, blocking and shadowing, alignment and tracking and adverse atmospheric condition [9]. But among these challenges, the impact of atmospheric turbulence goes a long way to affect telecommunication link availability especially when it comes to long range communication systems is of major concern [10]. Various mitigation techniques are employed either at physical layer or network layer to mitigate the FSO challenges highlighted above. Mitigation techniques utilized at physical layer are: transmission of multiple beam, increased receiver field of view, adaptive optics, transmission relay, hybrid RF/FSO etc., while in network layer techniques used are: Re-transmission, Reconfiguration and Re-routing, control of Quality of Service and DNT, replay.

The paper is organized as follows: section II will describe fundamental of optical wireless communication, section III elaboration of free space optical channel model, section IV detailed effects of atmospheric attenuation on optical wireless communication. Then, section V discusses the mitigation techniques for FSO communication challenges. Last but not least, conclusion is presented in section VI by highlighting the future research ideas and benefit of implementing FSO system.

II. FSO SYSTEM ARCHITECTURE

Figure 1below provides the generalized free space optical system architecture.

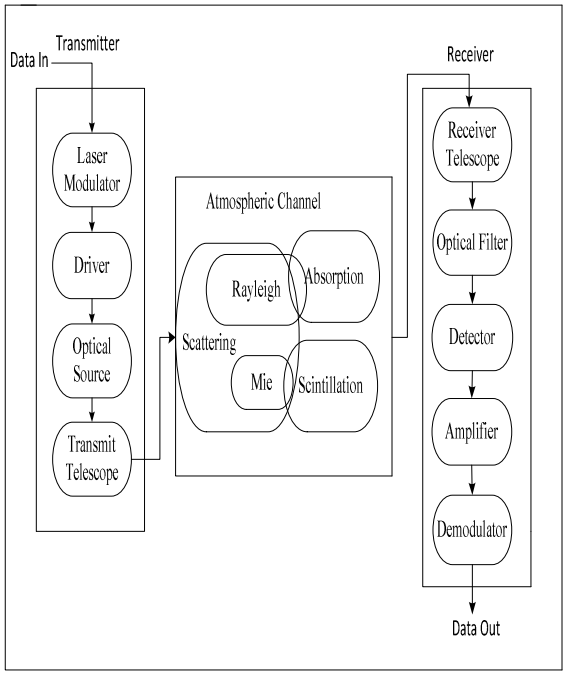


Fig. 1. Block diagram of FSO Fundamental Systems

A. *Transmitter*

The transmitter role in FSO is to change the electrical signal to a relate optical signal. There are four parts of transmitter, namely: laser modulation beam, circuit driver, source optical and transmitter's telescope. Optical sources commonly employed in FSO systems include vertical cavity surface, Fabry-Perot lasers, Distributed-feedback lasers, LED transmitting at wavelengths 850 nm, ~1300 nm, ~1500 nm, and near Infrared respectively. These sources are reliable and can ensure data rates up to 40 Gbps.

B. Atmospheric Channel

In free space optic links, atmosphere is the transmission medium. The atmosphere channel could be considered as a series of concentric gas layers round the earth. The troposphere, stratosphere and mesosphere are the three principal atmospheric layers outlined within the atmosphere. In FSO communication, the focus is more in the troposphere because it is where most weather phenomena occurs and FSO links operate at the lower part of troposphere layer.

The troposphere consists of several gases named as Nitrogen, Argon, Carbon dioxide, oxygen, water vapour. Also, particles are suspended in the air as addition to the tropospheric constituents and this type of suspension is termed as aerosols. The other phenomena are the rain, haze, fog and related weather condition which are seasonal based and location dependent. The cumulative effects of the tropospheric constituents often impede the optical signal that passes through them, hence leading to reduction in the power level of the signal strength receiving at the receiver. The significance of the refractive index of the troposphere under the influence of the temperature changes or inhomogeneity, lead to an atmospheric turbulence with noticeable results in the beam spreading, lack of coherency in the optical signal, polarization fluctuation and phase shifting.

C. Receiver

Receiver consists of the following parts, which are telescope receiver, optical filter, detector, amplifier and demodulator. Received signal is focused on the photo detector through the optical filter. The proportion of the aperture contributes to the noise background of the receiver; hence the filter does reduce the background noise. The detector converts the signal to back to electrical signal which further amplify and demodulated.

III. EFFECTS OF ATMOSPHERIC ATTENUATION ON OPTICAL WIRELESS COMMUNICATION

Despite the huge advantages of Optical wireless communication system, such as licence-free operation, unregulated huge bandwidth, low-cost front-ends, high security, etc., there are several challenges that come with the use of the technology and which need to be overcome in order to harness the full potential of the technology. The reasons for such weakening in the channel are due to different environmental factors such as mists, snow, haze rain, murkiness and so forth [11]. The optical beam experiences the following various losses while spreading through the environmental optical channel:

A. Atmospheric Turbulence

The random fluctuation of the received signal results from the inhomogeneity in temperature and pressure causes atmospheric turbulence. This turbulence degrades the performance of FSO communication systems [12]. The random variations in the temperature and the pressure of the atmospheric region through which the FSO signal has to pass causes the atmospheric turbulence. As a result of this, it redistributes the signal energy which leads to fluctuation in both the amplitude and phases of the propagation beam intensity at the receiver and significantly degrades the system performance [13, 14]. The refractive index of the air changes as the results of temperature changes in the atmosphere which causes the light beam to deviate from its intended path towards the receiver [15].

Kolmogorov Theory of turbulence, defined the outer scale as the largest cell size before the energy is injected into a region while the inner scale is defined as the smallest cell size before energy is dissipated into heat [1, 14].

Practically the outer scale has insignificant effect on the turbulence since its value is approximated as $L_o \to \infty$ while inner scale has significant impact on the turbulence because of its large value, resulting in higher irradiance variance especially during strong turbulence [16,17]. Consequently, Scintillation Index (SI) is the mostly used parameter to estimate the fluctuation emanating from atmospheric turbulence over the FSO links and is characterized in the equation below [14]:

$$\sigma^2_{SI} = \frac{\mathrm{E}(I^2)}{(\mathrm{E}(I))^2} - 1 \qquad (1)$$

where I is the received optical intensity (irradiance) and the expected E (.) is denoted as the expected value. Furthermore, defining SI in terms of Rytov variance and under the assumption of plane wave, weak turbulent condition can be provided (2) as [18]:

$$\sigma_{SI}^{\ 2} = \sigma_R^{\ 2} = 1.23 C_n^{\ 2} k^{7/6} L^{11/6} \quad (2)$$

and for spherical plane :

$$\sigma_{SI}^{\ 2} = 0.4\sigma_R^{\ 2} = 0.5 C_n^2 k^{7/6} L^{11/6} \quad (3)$$

where L represent the link distance, $k = 2\pi/\lambda$ is the optical wavenumber and C_n^2 is the refractive index structure parameter.

The turbulence conditions according to Rytov variance values can be classified approximately as follows [19]: weak turbulence conditions as $\sigma_R^2 \le 0.3$ moderate to strong turbulence conditions $0.3 \le \sigma_R^2 \le 5$ and saturation turbulence conditions $\sigma_R^2 \gg 1$. On the other hand the refractive index structure parameter C_n^2 measures the strength of the turbulence and its value depends on the altitude and link distance. It has large value at lower altitude due to heat transfer between the air and the earth surface [14]. [20] anticipated that for homogenous turbulence, the C_n^2 is independent on the distance especially for terrestrial FSO systems. The table 1 below represent the range values of refractive index structure parameter and their turbulence values:

TABLE I. RANGE VALUES OF REFRACTIVE INDEX AND THEIR TURBULENCE VALUES

Range value	Turbulence value
From $10^{-12} m^{-2/3}$	Strong turbulence
To $10^{-17} m^{-2/3}$	Weak turbulence

Authors in a literature proposed several empirical models of C_n^2 based on experimental measurement in different locations, time of the day, wind speeds, terrain type and so on, for the estimation of turbulent strength, the most widely used empirical model of Hufnagle-Valley (H-V) is utilized to characterize C_n^2 and is provided by equation below [12, 14, 21]:

$$C_n^{\ 2} = 0.000594 \left(\frac{v}{27}\right)^2 (10^{-5} h)^{10} \exp\left(-\frac{h}{1000}\right) + 2.7\times10^{-16} \exp\left(-\frac{h}{1500}\right) A \exp\left(-\frac{h}{1000}\right) \quad (4)$$

where h represent the altitude value, v the root mean square (RMS) of the wind speed, and A is the normal value of $C_n^2(0)$ at ground level.

1. Atmospheric Turbulence Distributions

Atmospheric turbulence induced-fading in FSO systems is one of the most substantial channel impairments that degrade the system performance. It is extremely challenging to definite the exact PDF that fit the statistics of optical radiation received through the atmospheric turbulence. It is probably impossible to study and predict atmospheric turbulence conditions over FSO links as there are no valid universal statistical model due to mathematical complexity involved in its modelling. Due to this challenge several authors have proposed different atmospheric models based on the turbulence regimes and below are the mostly accepted two models:

1.1 Lognormal Turbulence Model

Lognormal turbulence model is a suitable statistical distribution to emulate the irradiance fluctuations at the receiver's side due to the scintillation effect for the cases of weak turbulence conditions [19]. The model is derived from the Rytov approximation theory which require the magnitude of the scattered field wave to be small, compared to the unperturbed phase gradient. The corresponding PDF for h is provided by eqtion (5) [14, 19]:

$$f_h(h) = \frac{1}{h\left(\sqrt{2\Pi\sigma_h^2}\right)} \exp\left(\frac{\left(\ln(h) + 0.5\sigma_h^2\right)^2}{2\sigma_h^2}\right) \quad (5)$$

where σ_h^2 is the log-amplitude variance, and $\sigma_h^2 \approx \sigma_R^2/4 = 0.31\, C_n^2 k^{7/6} L^{11/6}$. This depends on the channel's characteristics and provided by this equation as shown below [3, 22, 23]:

$$\sigma^2 = \exp\left[\frac{0.49\sigma_i^2}{\left(1+0.65d^2+1.1\sigma_i^{12/5}\right)^{7/6}} + \frac{0.5\sigma_n^2\left(1+0.69\sigma_i^{12/5}\right)^{-5/6}}{\left(1+0.9d^2+0.62d^2\sigma_i^{12/5}\right)^{5/6}}\right] - 1 \quad (6)$$

where $d = 0.5 D_R \sqrt{k/L}$ [18, 22]

1.2 Gamma-Gamma Turbulence Model

Gamma-gamma model is another important model for multiplicative random process which was proposed by Andrew et al as a valid PDF for modelling FSO links over a wide range of turbulence conditions from weak to strong because it offers an excellent fit with experimental data while [18, 20, 24, 25, 26] stated that this model proved to be accurate in describing the irradiance fluctuations due to scintillation effect for moderate to strong turbulence channels.

Thus, the PDF for the random variable h can be expressed as [18, 26]:

$$f_h(h)=\frac{2(\alpha\beta)^{\frac{\alpha+\beta}{2}}}{\Gamma(\alpha)\Gamma(\beta)}h^{\frac{\alpha+\beta}{2}-1}K_{\alpha-\beta}\left(2\sqrt{\alpha\beta h}\right),\ h>0,\alpha>0,\beta>0 \quad (7)$$

where $\Gamma(\bullet)$ and $K_{\alpha-\beta}$ are defined as gamma function and α and β are scintillation parameters which are defined as the effective number of large scale and small scale eddies respectively, and is specified by the equation below [13, 14]:

$$\alpha=\left[\exp\left(\frac{0.49\sigma_R^2}{\left(1+1.11\sigma_R^{12/5}\right)^{5/6}}\right)-1\right]^{-1} \quad (8)$$

$$\beta=\left[\exp\left(\frac{0.51\sigma_R^2}{\left(1+0.69\sigma_R^{12/5}\right)^{5/6}}\right)-1\right]^{-1} \quad (9)$$

where σ_R^2 is the Rytov variance, and the SI to quantify the level of turbulence can be expressed as [14, 27]:

$$\sigma_{SI}^2=\frac{1}{\alpha}+\frac{1}{\beta}+\frac{1}{\alpha\beta} \quad (10)$$

The Gamma-Gamma turbulence PDF will degenerate to another PDF called K-distribution when β is assigned to unity, which is extensively utilized to model strong turbulence condition. The accuracy of this PDF model under strong turbulence conditions was validated through Experimental data used by Andrew and Philip [26]. At times over a distance of approximate 1km or when the scintillation index is restricted to the range of (2,3) this model provides accurate results [14] while when $\beta=1$ the K-distribution can be stated as:

$$f_h(h)=\frac{2(\alpha\beta)^{\frac{\alpha+1}{2}}}{\Gamma(\alpha)}h^{\frac{\alpha}{2}-1}K_{\alpha-\beta}\left(2\sqrt{\alpha h}\right),\quad h>0 \quad (11)$$

Also the SI, for K-distribution can be obtained as:

$$\sigma_{SI}^2=\frac{2+\alpha}{\alpha} \quad (12)$$

B. Atmospheric Loss

Absorption, scattering and attenuation are caused by the interaction of the molecules and particles with light in the atmosphere. The optical propagation signal quality can be affected by these interactions of atmospheric attenuations. Fog is the most important factor among the various atmospheric attenuation that have an effect on optical signal, this due to the size of fog particles which are comparable to the transmission wavelength of optical and near infrared waves.

This avoids the episode light from its underlying course, causing a spatial, precise and transient spread [25].

IV. Mitigation Techniques for FSO Communication Challenges

Debasement in the nature of received signal caused by atmospheric channel deteriorate the BER execution of the FSO framework. Various types of mitigation techniques are utilised to advance free space optic system reliability for atmospheric turbulences. Within either physical or network layer this mitigation procedures can be utilized. Mitigation techniques used in physical layer are transmission of multiple beam which increase FOV receiver, AO (adaptive optics), transmission relay, hybrid FSO etc. to enhance the enactment and the accessibility of free space optics framework in network layer, the following mitigation techniques are utilised: Packet retransmission applied either in network, re-routing or FSO link and quality of service control (QoS) and re-play of data, etc. Table 2 below provides various mitigation techniques employed in FSO communication.

Table 2: Turbulence Mitigation Techniques in FSO

Mitigation Techniques	
Physical layer methods	**TCP upper layer methods**
1. Aperture Averaging 2. Diversity 3. Adaptive Optics 4. Modulation and Coding 5. Background Noise Reflection 6. Hybrid FSO link	1. Re-Transmission 2. Re-routing and reconfiguration 3. Quality of service control 4. DTN, Replay

V. Conclusion

Reecently there have been remarkable increase of multimedia users and the internet traffic, this have acquired a significant strain on Radio Frequency system operating at lower data rates. Recently FSO proved to be an important replacement to radio frequency as it provides line of sight connection between remotes sites with very high bandwidths. Besides this merits of the FSO it suffers some drawback of atmospheric turbulence (geometric loss, atmospheric loss, and ambient noise) as it depend on the propagation of the laser beam through the atmosphere. This turbulence leads to change in the refractive index due to the fluctuations in the density of the air causing misalignment loss (due to the narrowness of the transmitted beam just a few centimeters movement can cause a large misalignment between the transmitter and the receiver which interrupts the communication link. Diversity, adaptive optics, error control codes, modulations, etc are few blurring mitigation strategies that were proposed for both Radio Frequency and Free Space Optics. Unmistakably this innovation appears to bring high development prospects in the near future.

Future scope of this paper is to develop a intelligence based technique that can dynamically adjust the system's signal that will enable the service providers to provide the optimum QoS signal at all times at less cost dispite different levels of signal attenuation.

ACKNOWLEDGMENT

I would like to acknowledge Tshwane University of Technology and my supervisors Prof P. Owolawi, Prof S. Ojo and Prof MI Mphahlele for the opportunity afforded to me to register for my Doctoral degree and their guidance in this research.

REFERENCES

[1] M. A. Khalighi and M. Uysal, "Survey on free space optical communication: A communication theory perspective," IEEE Communications Surveys & Tutorials, vol. 16, no. 4, pp. 2231-2258, November 2014.

[2] A.A. Huurdeman, "Optical Telegraphy," in The Worldwide History of Telecommunications, IEEE, 2003, pp. doi: 10.1002/0471722243.ch3

[3] M.N.O. Sadiku and S.M Musa, "Free space communication; an overview". European scientific journal, vol. 12, No 9. 2016

[4] Z. Ghassemlooy, W. Popoola, and S. Rajbhandari, Optical wireless communications: system and channel modelling with MATLAB. Boca Raton, FL, USA: CRC Pres, 2013.

[5] T. Rappaport, S. Sun and R. Mayzus, 'Millimeter wave mobile communications for 5G cellular: it will work!', IEEE Access, 2013, 1, pp. 335–349

[6] F. Finnell, F. Http://searchtelecom.techtarget.com. Accessed 19 December, 2016.

[7] Http://RFdesign.com. "Multigigabit wireless technology at 70 GHz, 80 GHz and 90 GHz," RF design. 2006

[8] Http://4gon.co.uk. Accessed online on 11 April, 2017

[9] V.W.S. Chan, "Free-space Optical Communications," J. Lightwave Tech., vol. 24, no. 12, pp. 4750–4762. 2006

[10] N.S. Kopeika and J. Bordogna, "Background noise in optical communication systems", Proc. IEEE, 1970, 58, (10), pp. 1571–1577

[11] Henniger and O. Wilfert, "An Introduction to Free-space Optical Communications," *Radioengineering,* vol. 19, no. 2, June 2010.

[12] E. J. Lee and V. W. Chan, "Part 1: Optical communication over the clear turbulent atmospheric channel using diversity," *IEEE journal on selected areas in communications,* vol. 22, no. 9, pp. 1896-1906, November 2004.

[13] J. H. Churnside, "Aperture averaging of optical scintillations in the turbulent atmosphere," Applied Optics, vol. 30, no. 15, pp. 1982-1994, May 1991.

[14] K. Kiasaleh, "Performance of coherent DPSK free-space optical communication systems in K-distributed turbulence," IEEE transactions on communications, vol. 54, no. 4, pp. 604-607, April 2006.

[15] X. Zhu and J. M. Kahn, "Free-space optical communication through atmospheric turbulence channels," *IEEE Transactions on communications,* vol. 50, no. 8, pp. 1293-1300, August 2002.

[16] L. Yang, X. Song, J. Cheng, and J. F. Holzman, "Free-Space Optical Communications Over Lognormal Fading Channels Using OOK With Finite Extinction Ratios," IEEE Access, vol. 4, pp. 574-584, 2016.

[17] V. I. Tatarski, "Wave propagation in a turbulent medium," Courier Dover Publications, 2016.

[18] L. C. Andrews, R. L. Phillips, and C. Y. Hopen, "Laser beam scintillation with applications" vol. 99: SPIE press, 2001.

[19] K. Kiasaleh, "Performance of coherent DPSK free-space optical communication systems in K-distributed turbulence," IEEE transactions on communications, vol. 54, no. 4, pp. 604-607, April 2006.

[20] A. Consortini, J. H. Churnside, R. J. Hill, and F. Cochetti, "Inner-scale effect on irradiance variance measured for weak-to-strong atmospheric scintillation," Journal of Optical Society of America, vol. 10, no. 11, pp. 2354-2362, November 1993.

[21] E. Jakeman, "On the statistics of K-distributed noise," Journal of Physics A: Mathematical and General, vol. 13, no. 1, p. 31, 1980.

[22] L. C. Andrews, R. L. Phillips, C. Y. Hopen, and M. Al-Habash, "Theory of optical scintillation," Journal of the Optical Society of America vol. 16, no. 6, pp. 1417-1429, June 1999.

[23] L. C. Andrews and R. L. Phillips, "Laser beam propagation through random media" vol. 152: SPIE press Bellingham, WA, 2005.

[24] A. García-Zambrana, B. Castillo-Vázquez, and C. Castillo-Vázquez, "Asymptotic error-rate analysis of FSO links using transmit laser selection over gamma-gamma atmospheric turbulence channels with pointing errors," Optics express, vol. 20, no. 3, pp. 2096-2109, January 2012.

[25] M. Al-Habash, L. C. Andrews, and R. L. Phillips, "Mathematical model for the irradiance probability density function of a laser beam propagating through turbulent media," Optical Engineering, vol. 40, no. 8, pp. 1554-1562, August 2001.

[26] M.-A. Khalighi, N. Schwartz, N. Aitamer, and S. Bourennane, "Fading reduction by aperture averaging and spatial diversity in optical wireless systems," *Journal of optical communications and networking,* vol. 1, no. 6, pp. 580-593, November 2009.

[27] L. Yang, X. Song, J. Cheng, and J. F. Holzman, "Free-Space Optical Communications Over Lognormal Fading Channels Using OOK With Finite Extinction Ratios," *IEEE Access,* vol. 4, pp. 574-584, 2016.

Innovative Method for Power Factor Correction using a Solar Plant as a Source of Reactive Power

Esrom Mahlatsi Malatji
Dept. of Electrical and Electronic Engineering Technology
University of Johannesburg
South Africa
emalatji@uj.ac.za

Billy Chabangu
Dept. of Electrical and Electronic Engineering Technology
University of Johannesburg
South Africa
billychabangu93@gmail.com

***Abstract*—Due to the low power factor of an electrical system there is a high demand of reactive power from the transformers which results in overloading of the transformers. Traditionally, this problem is solved by adding a capacitor in parallel with the load in order to correct the power factor. In this paper a new innovative method which utilizes the solar plant at night when it is idling or during the day when it is not producing real power. During these times a solar plant is used as a source of reactive power, the inverters used in solar plants are capable of being used as a source of reactive power. The case study used to test this innovative method is a food processing plant with multiple transformers. The results show that it is possible to use a solar plant for power factor correction and hence relieve the transformers from overloading issues.**

Keywords—transformer, solar, power factor, reactive power, overloading

I. INTRODUCTION

Transformers play an important role in electrical networks because they transform the current from a high voltage to a low voltage which can be handled by electrical equipment in industries and also domestically. When a higher load is drawn from the transformers then a phenomenon called overloading occurs. This places great strain on transformers and eventually reduces the life span of the transformer. This is also not safe as this can result in transformers catching fire and resulting in a great fire hazard because of the high temperatures produced with drawing high currents.

There are many ways to address the challenge of overloading such as replacing a smaller transformer with a bigger one. But the cost of upgrading a transformer can be quite excessive and other methods should be considered. One study considered methods to manage the overloading of the transformer beyond its name plate rating and the study found out that the limiting factor of overloading capability was the ancillary equipment [1]. Other studies [2] – [5] concluded that improving the cooling system of the transformer can improve the overloading capability of the transformer.

Having a system with bad power factor can result in an unnecessary demand on the transformer, which might lead to overloading of the transformer and other equipment in a power system network. Introducing reactive power in a system is one way to reduce the overloading of the transformer. The most common way of introducing reactive power in the system is to make use of capacitor banks which is the technology that has been highly used in the past. Other studies show the application of capacitors for reactive power control in industrial networks [6] – [8]. The studies in [9] – [11] show a new method of generating or supplying reactive power to a system. The studies suggest that a solar plant can be used as a source of reactive power during the night when the plant is not producing any real power. This is made possible by the fact that solar plants use Voltage Source Inverters (VSC) to convert Direct Current (DC) power to Alternatiig Current (AC). The same component is at the heart of STACOM which is a Flexible AC Transmission System (FACTS) device that can give or absorb reactive power. The VSC uses capacitors that are charged as a DC source and this in turn produces an AC voltage which is in synchronism with the AC system [12]. This means with a little modification a solar plant can be converted to STATCOM at night and during the day to some extent.

One of the first few studies to propose the concept of utilizing a solar plant as a STATCOM in order to regulate the grid voltage is found in [13]. The concept was proposed for distribution feeders and a study was conducted in MATLAB in order to validate the concept. The concept was then applied to a wind farm which has high output variations, and this proved useful as the PV plant could assist in voltage regulation even in this case.

The research in [14] conducted a case study to prove the concept of using a solar plant as a STATCOM with the application being voltage regulation and reactive power compensation which would take place at the point of common coupling. The benefits of this was the increase in transfer limits and improved stability. The research in [15] demonstrated that solar technology is rather expensive and often remains idle due to a lack of sunlight on some days. In this research a simple control algorithm was used which altered a PV plant to operate as a STATCOM. The results were simulated in MATLAB and were found to be effective especially in the case of reducing voltage sags. In [16] an illustration of a comprehensive plan for testing, installation and validation of the PV plant as a STATCOM is given. The installation took place at headquarters of London Hydro Distribution Company. This company is based in Canada and has a 10 kW solar rooftop. The results showed an improvement in the electrical performance of the system and also proved to be an extra source of revenue. The Blue Power Corporation in [17] also showcased their optimal scheduling of PV plant when operating as a STATCOM and the controller which is simulated using PSCAD, is developed in the university laboratory. In this research the transient and steady state performance were tested, and the results were satisfactory.

The research in this paper proposes the use of a solar plant as a source of reactive power to compensate the system which will help with relieving the overloaded transformers and the power factor correction. The contribution of this research is to show that a solar plant can be used as a source of reactive power. This means it can be used as power factor correction equipment instead of capacitors. The advantage of using the solar plant is that one gets to increase the utilization of this expensive asset which is idling almost 50% of the time.

The rest of the paper is as follows; Section 2 outlines the overloading challenge of the transformer, then the innovative solution is given in Section 3. This is followed by the results and discussion in Section 4 and the paper is concluded in Section 5.

II. TRANSFORMER OVERLOADING PROBLEM

The design of a solar system that will help with the challenge of overloading experienced by the transformer, is considered in this section. The solar system will relieve the transformer from overloading in two ways. Firstly, by providing real power to the network system during the day. This will remove some load from the transformers. Secondly by applying innovative methods which convert the solar system into a reactive power source during the night when the system is idling and also during the day when the solar system is not producing real power. By improving the reactive power this will improve the power factor of the system and reduce the loading on the transformers.

The case study considered in this paper is a food processing plant which has 7 transformers witch experience overloading challenges. A logger has been installed to record the load of the transformers which operate for 24 hours a day. The data was logged on an hourly basis for a period of 5 weekdays. The reason for this is that the plant operates roughly the same each and every week. The information about the transformers are given in Table 1. The current given in Table 1 is the highest value that was recorded during the course of the week. The power factor of the system is 0.88 and the desired power factor is 0.99. The current measurements for each transformer are given in Figure 1.

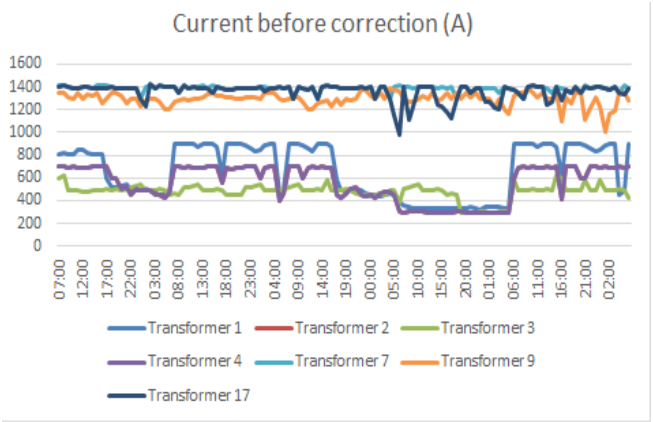


Fig 1: The current measurement of the transformers before power factor correction

TABLE I. TRANSFORMER INFORMATION

	Location	Current	Transformer Size	Power factor
Transformer 1	Supply Bulk processing	909A	600 KVA	0.88
Transformer 2	Bus coupled with 3 & 4 Supply Frick compressors and boosters (Tunnels)	722A	500 KVA	0.88
Transformer 3	Bus coupled with 2 & 4 Supply Frick compressors and boosters (Tunnels)	722A	500 KVA	0.88
Transformer 4	Bus coupled with 2 & 3 Supply Frick compressors and boosters (Tunnels)	722A	500 KVA	0.88
Transformer 7	Supply QA lab, Sullair compressors	1443A	1000 KVA	0.88
Transformer 9	Supply Condensers and battery bay	1443A	1000 KVA	0.88
Transformer 17	Supply Mycom and Sullair compressors	1443A	1000 KVA	0.88

From the information given in Table 1 there are a lot of parameters that must be calculated in order to adequately size the solar system so that the power factor can be increased from 0.88 to 0.99. The same procedure would be followed if there was a need to calculate the size of capacitors needed to move the power factor to 0.99. Those parameters were calculated, and they are given in Table 2. These parameters are only for the peak values, but the graph will be plotted using the data from the entire data for the week. The parameter that needs to be determined is the actual power consumption and this can be calculated using equation (1). The required reactive power, which is the reactive power that will move the power factor from 0.88 to 0.99, is calculated using equation (2) and it is also given in Figure 2. Figure 2 captures the current five-day profile. The corrected current is the current being pulled from the transformers after the power factor correction is calculated by making use of equation (3). The corrected load is the load of the transformers after the power factor correction and this is given by equation (4). The current reduced and load reduced are the amount of current and load reduced because of the influence of the power factor correction. These are given by equation (5) and (6) respectively.

Calculate the actual power

$$P_{Trfr,t} = \sqrt{3} \times V \times I_{Trfr,t} \times cos\theta_1 \quad (1)$$

$P_{Trfr,t}$ is the real power of the transformer with the current given by $I_{Trfr,t}$. The voltage of the system is V while the power factor before correction is θ_1.

Calculate the required power factor correction (kVAr)

The reactive power required to move the power factor from θ_1 to θ_2, where θ_2 is the power factor after correction, is given by equation (2).

$$R_{kVAr} = P_{Trfr,t}(\tan(\cos^{-1}\theta_1) - \tan(\cos^{-1}\theta_2) \quad (2)$$

The reactive power should be supplied by the solar plant when it is not producing real power which is mainly at night or on days were there is no sun. From Table 2 it is seen that the required reactive power is given by equation (3), as the Total Kilovar Required (TKrR).

$$\text{TkrR} = 220 + 175 + 175 + 175 + 482 + 482 + 482 \quad (3)$$
$$= 2191 \text{ kVar}$$

which is calculated from individual peak values. Figure 2 shows the reactive power that should be supplied by the solar plant.

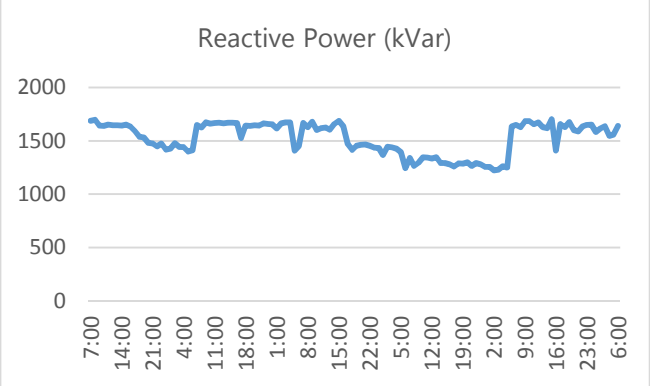


Fig 2: Required reactive power

Calculating the corrcted amps

$$I_{Trfr,t}^{corrected} = \frac{P_{Trfr,t}}{\sqrt{3} \times V \times cos\theta_2} \quad (4)$$

Corrected load

$$L_{Trfr,t}^{corrected} = \sqrt{3} \times V \times I_{Trfr,t}^{corrected} \quad (5)$$

Reduction in amps

$$RedA_{Trfr,t} = I_{Trfr,t} - I_{Trfr,t}^{corrected} \quad (6)$$

Reduction in Load

$$RedL_{Trfr,t} = L_{Trfr,t} - L_{Trfr,t}^{corrected} \quad (7)$$

From Table 2 it can be confirmed that increasing the power factor from 0.88 to 0.99 there is a current which can be reduced. This implies having another source that can add the required amount of current or power that will relieve the transformers, this will achieve the same objective. The real power that would be needed in the research is supplied by a solar plant.

TABLE II. PEAK VALUES

	Actual Power (kW)	Required reactive power (kVAr)	Corrected Current (A)	Corrected Load (KVA)	Current reduced (A)	Load Reduced (KVA)
TRF 1	554.2	220.157	808	560	101	40
TRF 2	440.2	174.87	642	445	80	55
TRF 3	440.2	174.87	642	445	80	55
TRF 4	440.2	174.87	642	445	80	55
TRF 7	879.7	481.693	1283	890	160	110
TRF 9	879.7	481.693	1283	890	160	110
TRF 17	879.7	481.693	1283	890	160	110

The solar plant that needs to be installed to achieve the required results without installing power factor correction capacitors is given by equation (8) as the Total Kilowatt Required (TKR).

$$\text{TKR} = 58 + 46 + 46 + 93 + 93 + 93 + 93 \quad (8)$$
$$= 522 \text{ kW}$$

The power was calculated using the reduced currents from Table 2. The profile of the required real power from the solar plant is given in Figure 3.

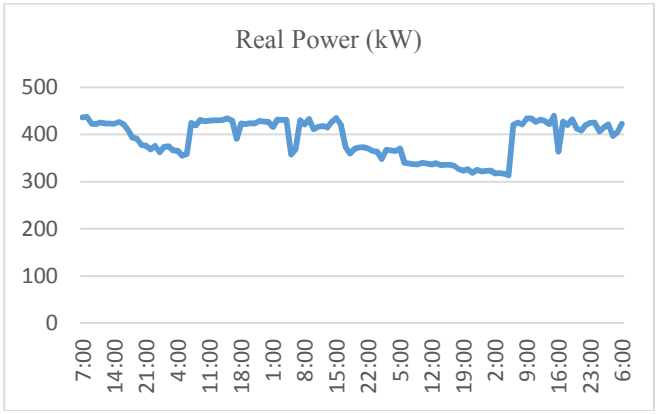


Fig. 3: Required real power

III. INNOVATIVE SOLUTION

To be able to supply both the reactive and real power of the profiles given in Figure 2 and 3, a solar plant given in Figure 4 was used. A 2000 kW plant must be designed in order to supply the required real and reactive power. Figure 5 depicts the solar plant operating in real and reactive power producing modes. Equal power is produced in both modes which is 2000 kW and 2000 kVar.

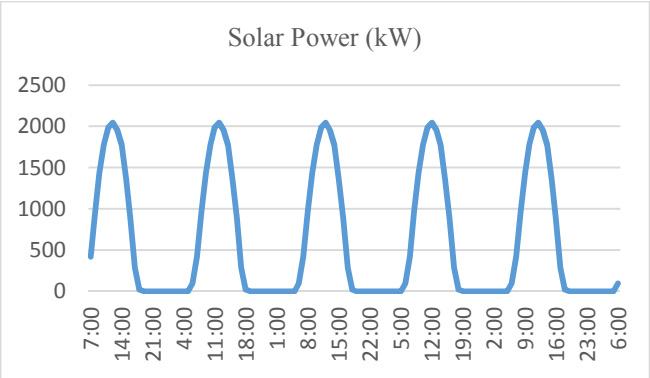


Fig. 4: Output solar plant

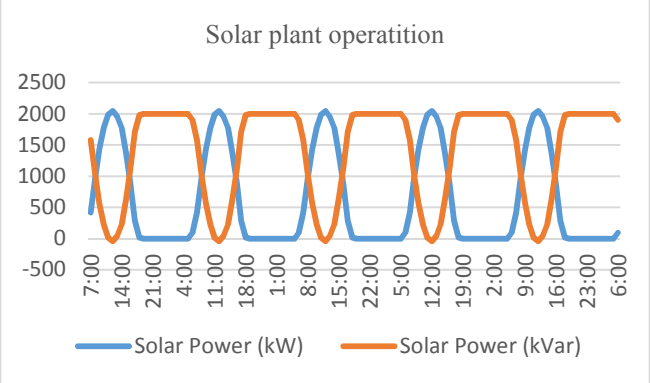


Fig. 5: The operation of the solar in real and reactive modes

Figure 6 depicts the required reactive power and the reactive power that would be supplied by the solar plant. The Figure shows that there are periods where there is more reactive power and periods when there is insufficient reactive power. The same principle is depicted for real power as shown in Figure 7.

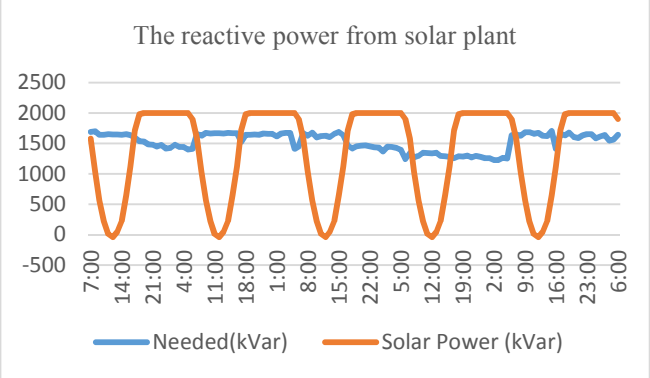


Fig. 6: Reactive power compensation

Figure 8 shows the results obtained after the power factor correction was applied. This shows the reduction in the current that is being drawn from the transformers, this can be seen when you compare the output of Figure 1 and Figure 2. These results were obtained using Microsoft excel. This profile in Figure 8 is achieved by either injecting reactive power from the solar plant which will be mainly during the night or the real power which will be during the day when the solar resources are available.

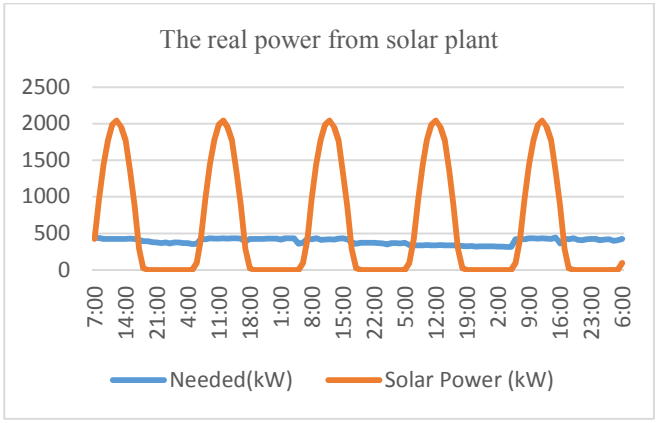


Fig. 7: Real power compensation

IV. Result and Discussion

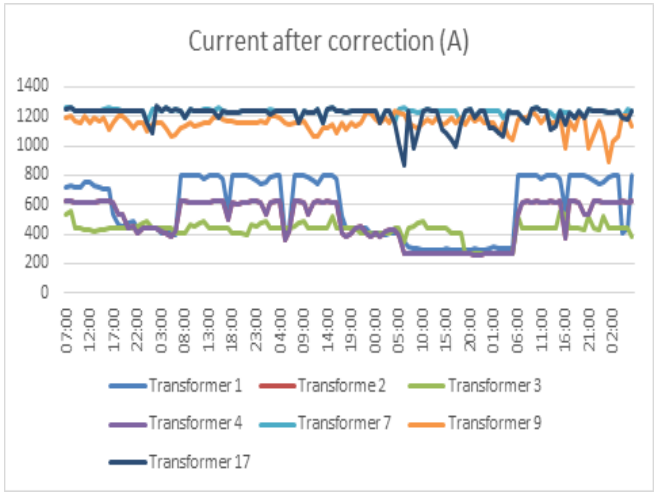


Fig 8: The currents of the transformers before power factor correction

It can be seen from Figure 2 and Figure 3 that the amount of reactive power and real power are considerably unequal. This presents a challenge because it suggests that there is a need to build a large solar plant to accommodate the high requirements of the reactive power. This is the reason that a 2 MW solar plant is designed as shown in Figure 4.

V. Conclusions

In this reaseach it is shown that a solar plant can be used to perfom power factor correction in order to relieve the transformers in serving the requirement of the system. The important aspect of using a solar plant for power factor correction is that it can work in both reactive and real power modes. The other advantage of this method is that when the plant is producing real power it can also save a lot of money for the plant. That's because the plant would not be buying some of the power from the grid. The greatest disadvatage of using a solar plant for this purpose is that when reactive power is used, more capacity of the plant is needed compared to when it is operating in the real power mode. To solve this issue a different kind of inverter/converter should be designed to provide more reactive power than the real power.

REFERENCES

[1] A. Pasricha, Crow, M.L. "A method of improving transformer overloading beyond nameplate rating." North American Power Symposium (NAPS), Charlotte, NC, USA, 2015.

[2] J. G. Hwang, M. Zahn, F. M. O. Sullivan, L. A. Pettersson, O. Hjortstam, and R. Liu, "Effects of nanoparticle charging on streamer development in transformer oil-based nanofluids," Journal of Applied Physics, vol. 107, no. 1, 2010.

[3] S. C. Kaishian, D. C. Pixler, T. A. Roebke, S. D. Day, and W. K. Siebert, "Cooling of electrical components with split-flow closed-end devices," May 2007, US Patent 7, 212,406.

[4] E. Sines, "Electrical power cooling technique," 2001, US Patent 6, 259, 347.

[5] R. S. Downing and G. I. Rozman, "Planar magnetics with integrated cooling," August 2001, US Patent 6, 278, 353.

[6] L. Gebhardt and B. Richter, "Surge arrester application of MV Capacitor banks to mitigate problems of switching restrikes", Int. Conf. on Electricity Distribution, Vienna, 2007.

[7] T. Kamakawaji, T. Shioiri, T. Funahashi, Y. Satoh, E. Kaneko and I. Oshima, "An investigation into major factors in shunt capacitor switching performances by vacuum CBs with copper chromium contacts". IEEE Trans Power Deliv, vol. 8, no. 4, pp. 1789–95, 2014

[8] D. Koch, Control equipment for MV capacitor banks, ECT142. Grenoble: Schneider-Electric; 1992.

[9] M. S. Eswar Rao and K. Obulreddy, "Utilization of PV solar Farm as Statcom during night hours in a distribution Network". International Journal of Science, vol. 3, no. 2, pp.33-39, 2014.

[10] R. Varma, S. A. Rahman and T. Vanderheide, "New Control of PV Solar Farm as STATCOM (PV-STATCOM) for Increasing Grid Power Transmission Limits during Night and Day," IEEE Trans Power Deliv . vol. 30, no. 2, pp.755-763, 2015.

[11] M. Zuercher-Martinson, "Smart PV inverter Benefits for Utilities," Renewable Energy WORLD [online]. [Accessed August 08, 2017]. Available from: <www.lgenergy.com>.

[12] R. M. Mathur and R. K.Varma, "Thyristor- Based FACTS controllers for Electrical Transmission systems," Newyork: Wiley- IEEE press, 2002.

[13] R. K. Varma, V. Khadkikar and R. Seethapathy, "Night-Time Application of PV Solar Farm as STATCOM to Regulate Grid Voltage", IEEE Transactions on Energy Conversion (Letters), vol. 24, no. 4, pp. 983-985, 2009.

[14] A. Surendran and A. S Mohan, "A case study on PV STATCOM with different controls for increasing grid power transmission limits during night and day" International Journal of Scientific & Engineering Research, vol. 7, no. 4, pp. 487-492, 2016.

[15] N. L Prasanthi and M.S Krishnarayalu, "Photovoltaic solar plant as a statcom during dark periods in a distribution network" International Journal of Engineering Research & Technology (IJERT), vol. 1, no. 9, pp. 1-7, 2012.

[16] R. K. Varma, S.. S. Rangarajan, I. Axente and V. Sharma, "Novel application of a PV Solar Plant as STATCOM during Night and Day in a Distribution Utility Network", IEEE PES Power Systems Conference & Expo, Phoenix, Arizona, USA, 2011.

[17] R.. K. Varma, B. Das, I. Axente, and T. Vanderheide,"Optimal 24-hr Utilization of a PV Solar System as STATCOM (PVSTATCOM) in a Distribution Network", Proc. IEEE PES General Meeting, Detroit, USA, 2011.

Climatic Effects on Free Space Optics Link: South African Climate

Solly P Maswikaneng
Dept. Information Technology
Tshwane University of Technology
Pretoria, South Africa
maswikanengps@tut.ac.za

Pius A Owolawi
Dept. Computer Systems
Tshwane Univesirty of Technology
Pretoria, South Africa
owolawipa@tut.ac.za

Sunday O Ojo
Dept. Computer Science
Tshwane University of Technology
Pretoria, South Africa
ojoso@tut.ac.za

Z Mahlobogwane
Dept.Computer Systems
Tshwane University of Technology
Pretoria, South africa
mahlobogwanelz@tut.ac.za

Maredi I Mphahlele
Dept. Dean of ICT
Tshwane University of Technology
Pretoria, South Africa
mphahlelemi@tut.ac.za

Abstract— **The growth in demand of high data rate, reliable quality of service, low cost technology and secured connectivity have been telecommunication users' concerns. Free Space Optics (FSO) is the best choice for meeting these demands as it offers data rates in Gbps with the most secure communication because light beam is prone to eavesdropping. However, usage of FSO as a communication link relies upon local weather condition. Atmospheric attenuation of fog, rain and snow are the most limitation of free space optic availability, as they lead to the diminution of limited available bandwidth. This paper investigated the link performance under climatic conditions such as fog, rain and snow by at various wavelengths of 650nm, 750nm, 850nm, 950nm and 1550nm. It is further deals with FSO performance targeting to acquire minimal Bit Error Rate (BER), suitable transmitting power, optimum link margin and reliable data transmission rate at maximum link distance of 1km. The studies observed that received power signal and the link distance have an inverse correlation because once the signal reaches to an extended distance; it faces greater climatic attenuation due to scattering and absorption lessening.**

Keywords—Free space optics, atmospheric attenuation, communication link, light beam, bit error rate

I. INTRODUCTION

The ever increasing demand of services consuming high data rates is increasing daily. Traditional microwave communication systems can no longer support this high bandwidth demand. There is demand of telecommunication technology that will fulfil high data rate demand with reliable quality of service at lowest cost as possible and yet provide a secured link between the users and FSO's systems. By taking all the facts into consideration, Free Space Optics (FSO) is consider to be the best choice to meet these demands. FSO offers data rates in Gbps with the most secure communication because light beam is prone to eavesdropping. Provides many advantages as compared to other wireless technologies such as free licensing, ease of installation, low capital equipment cost and very high data rates.

FSO has found fascinating applications in access networks (last mile solution), airborne & inter-satellite communication, disaster recovery, inter-building connections (point to point or multipoint) and short term installation for certain events, backhaul for cellular systems, broadband access to remote areas [1,2]. Despite of its several advantages, the availability of FSO is the main concern since this technology uses light propagation which is influenced by different weather conditions, hence attenuate transmitted signals. FSO utilizes open space channel as an interface for establishing link between transmitter and receiver. Therefore, weather conditions must be examined before practical implementation of FSO link. Absorption, scintillation and scattering are the factors, which affect the FSO link [3, 4]. Authors [2][5] indicated that wavelength's optical strength has a comparable size aerosols molecules such as fog, rain droplets etc.

The dominant of weather types depends on the geographical location. In the case of effect of haze, which is considered the prevailing component that influences the FSO link execution in various districts as the authors in [6], indicated. They further demonstrated that rain effects relying upon the fundamental aspect of lessening the Free Space Optics interface availability with a focus on the humid environment [7].

FSO link availability in this study is based on the visibility data collected from the South Africa Weather Services (SAWS). The data is processed, the FSO signal characterization is done, and performance analysis was implemented. This research study is arranged as follows: section II Critical atmospheric attenuations that affect the performance of Free Space Optics system, section III Evaluation performance of the free space optic system, IV Results and discussions and section V indicates Conclusion.

II. CRITICAL ATMOSPHERIC ATTENUATIONS THAT AFFECT THE PERFORMANCE OF FSO SYSTEM

FSO signal is weakened by the atmospheric condition as earlier mentioned, this resulted to transmitted signal being scattered or absorbed by the inherent molecular particles almost in the same size as the wavelength of the transmitted signal [8, 9]. Author [10] indicated that atmospheric conditions and the clouds cause the scattering of the pulse and the maximum serious climate situations, hence attenuates the transmitted signal. In the subsection of this paper, critical climatic conditions are explained as follows:

A. Rain Condition

The rain weakening of terrestrial FSO link relies on the quantity of rain, together with rain attenuation is in particular

excessive and significantly relied on diverse models of rain rate distribution with respect to FSO which operate at Terahertz wave bands [11].

Apart from rain rate distribution model, the consideration of raindrop size distribution play an important role in estimating rain attenuation along the FSO links. The work done by Marshal and Palmer proposes the use of raindrop size distribution model using a temperate data [12].

Attenuation of rain is wavelength independent. Therefore, a preferred form of exact connection between the precise rain constriction and rain rate is calculated by means of this power law:

$$A = kR^{\sigma} \tag{1}$$

where R represents the rain rate in mm/hr, while k and σ represent power law constants depends on the geographical region where the data is collected and they relies on frequency, temperature and microstructure of rain. In the work done by authors [13], they indicated that by using power law technique, the values k and σ are attained through utilising the curve fitting or factor equivalent systems.

This power law approach can be attained when the intensity of rain data is accessible and it indicates the forecast of optical diminution owing to rain. Table I below offers the constants values of the k and σ for particular rain diminution models of FSO as endorsed by ITU-R.

TABLE I: RAIN ATTENUATION PREDICATION MODEL PROPOSED BY ITU-R FOR FSO [14]

Model	Origin	k	σ	Note
Carbonneau	France	1.076	0.67	Temperature region
Japan	Japan	1.58	0.63	Temperature region

As indicated in the above Table I, the Charbonneau's approach proposed values to calculate the k and σ are grounded from the measurement done in France, the measurement was done in a region with a low rain rate as contrasted to higher rain rate in tropical region. Although, 1-minute rain rate integration is often considered, but many authors proved that there is a slight different in the 5-minute integration time rain rate [15]. It must be noted that later integration time rain rate is used in this paper. Since the Table I constants are based on temperate and South Africa being in a sub-tropical new values of 0.3988 for k and 0.7601[16] for σ are projected based on the measurement done and they are adopted in this study since South Africa falls under a subtropical region.

B. Attenuation of Snow condition

Snow are crystal ice available in extraordinary shapes and sizes [17]. Based on its characteristics, snow has a tendency of being larger than rain and fog [9]. More Snowfall rate have more impact on FSO transmitted signal than its counterparts do. Dry snow and wet snow are the two modes of snow characteristics. Author [18] indicated that a snowball of 8mm/hr, attenuated FSO signal at approximately 115dB/km may be located within 0.5km visibility range, 4 mm/hr for blizzard, attenuation of 38.667 dBm/km with visibility range of 1.5 km are observed [11, 17,18].

C. Attenuation of Fog

Author [9] indicated that the fog falls within the main catastrophic weather condition to FSO due to its composition of small water droplets with radii around the scale close to electromagnetic wavelengths. The particle length dissemination differs from distinctive tiers of fog [17].

The two major scattering mechanisms (the Rayleigh and Mie) are utilized in FSO. Author [18] differentiated these scatterings by considering Rayleigh scattering as preferable when the transmitted signal wavelength is greater than the size of the molecule in which it is transmitted while on the other hand Mie scattering happens when molecule estimate is equivalent to the wavelength-transmitted signal at a given wavelength. The equation for scattering transmittance is provided by:

$$\tau_s = e^{-\gamma_{fog} L} \tag{2}$$

where L is the length of the link and γfog represent attenuation due to fog is provided by [18]:

$$\gamma_{fog}(\lambda) = \frac{3.91}{V}\left(\frac{\lambda}{550}\right)^{-\sigma} \tag{3}$$

where V represents visibility in km, λ is wavelength (nm) and parameter σ is visibility dependence. Equation (3) indicates that the wavelength is inversely proportional to the attenuation for any weather conditions. For Kruse model σ is provided by:

$$\sigma = \begin{cases} 1.6 & V > 50\text{km} \\ 1.3 & 6\text{km} < V < 50\text{km} \\ 0.58V^{1/3} & V < 6\text{km} \end{cases} \tag{4}$$

In dense fog Kim model wavelength is free for low perceivability. In Kim model the expression of σ is provided by [19]:

$$\sigma = \begin{cases} 1.6 & V > 50\text{km} \\ 1.3 & 6\text{km} < V < 50\text{km} \\ 0.16V + 0.34 & 1\text{km} < V < 6\text{km} \\ V - 0.5 & 0.5\text{km} < V < 1\text{km} \\ 0 & V < 0.5\text{km} \end{cases} \tag{5}$$

It is observed that snow is the most adverse condition as compared to fog and rain.

III. RESULTS AND DISCUSSIONS

Simulation is done to highlight the impact of rain on FSO system and FSO links while making use of NRZ modulation techniques and PIN photodiode is utilized within the transmitting power source. Kruse model is chosen because of its wavelength reliance even as the estimation of the visibility data are taken into consideration in the eight stations in the South Africa for as long as 10 years. At the

same time, Rain data are also obtained for eight stations. These are integrated into the generated FSO signal to determine the attenuation contributions considering the two aforementioned weather factors. Table II indicates the parameter values used in the simulation.

TABLE II: PARAMETERS VALUES USED FOR SIMULATION

Operating parameters	**Value**
Transmitter Power (Pt)	0.005W
Laser Beam Divergence Angle	1 mrad
Transmitter Efficiency (τ_t)	0.9
Receiver Efficiency (τ_r)	0.9
Wavelength (λ)	850, 950 and 1550 nm
Range (L)	0.1 $\leq$ L $\leq$ 1 km
Visibility (V)	1 Km
Average Atmosphere Temperature	20 °C
Average Atmosphere relative Humidity (RH)	67.9%
Rain Rate (R_{rain})	18.3 mm/hr
Receiver Sensitivity (N_b)	0.00001 W
Data rate (R)	100*10^6 b/s
Boltzmann Constant (K)	$1.38x10^{-23}$ J.K
Absolute Photodiode Temperature (T_{PIN})	298 K
Dark Current (I_D)	10 nA
Responsibility (R_{PIN})	0.6 A/W
Electrical Bandwidth (B	0.5 GHz

Figure 1 represents the Durban power received at various weather conditions in dBm at the wavelength of 650nm against link distance. There is an inverse relationship between the acquired signal strength and the link distance because the relationship between the transmitted signal wavelength and the interacted atmospheric molecule wavelengths. The figure displays that there is high impact of attenuation during the heavy fog as compared to other conditions. Fog has a greater attenuation that have overall effect on the optical light propagation in various weather conditions.

TABLE III: SEASONAL OF PRECIPITATIONS IN SOUTH AFRICA

Station	Latitude	Longitude	Rain rate at 0.01%	Attenuation at 0.01%
Upington	-33.9881	22.453	29	5.129
Durban	-29.858	31.0292	53	8.132
Polokwane	-23.9045	29.469	77	10.809
East London	-33.015	27.912	43	6.980
Bloemfontein	-29.1	26.3	48	7.563
Emalahleni	-25.873	29.255	60	8.960
Pretoria	-25.731	28.218	50	7.848
Cape Town	-33.925	18.424	24	4.465

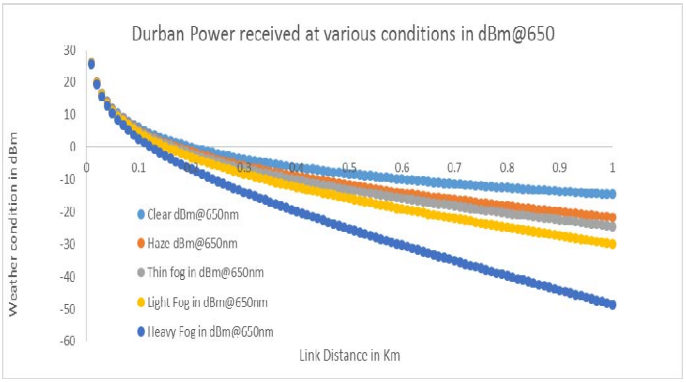


Fig. 1. Power received in different condition versus link distance

Rain Attenuation against Rain rate

Based on the equation (1), rain attenuation from eight stations in South Africa with varying rain rate is presented in Figure 2 below and is based on the new values of 0.3988 for k and 0.7601 are adopted for South Africa as a subtropical region. Table III below indicates the average rain rate exceeding 0.01% for different stations in South Africa. ITU-R indicated that the determination of the rain rate for 0.01% of the time comes from the fact that a good system must provide at least 99.99% reliability [18]. According to [18] in tropical region at microwave, and millimetre wave bands, rain is expected to be the major impairment at the bands mentioned while in the case of FSO, the effect is limited.

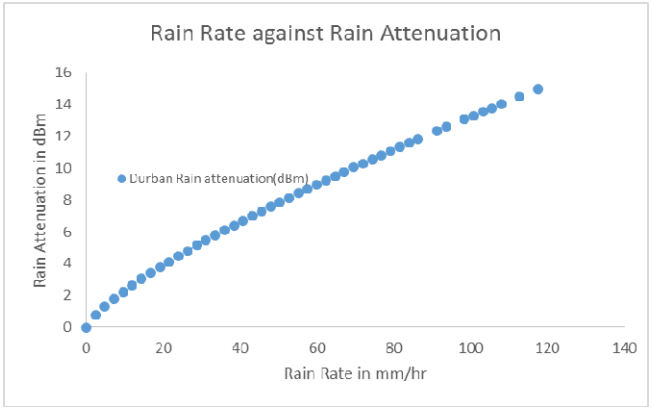


Fig. 2. Rain attenuation versus rain rate at Durban station

Figure 2 shows the atmospheric attenuation versus rainfall rate for Durban, other stations were not included in this paper as they have minimal impact as indicated in the table. When the rainfall rate increases, the effect of atmospheric attenuation on the FSO system increases slightly with minimal deviation. Therefore, the influence of attenuation on transmission of FSO systems is more prominent during heavy rainfall as compared to moderate and light rainfall.

Rain rate against percentage of exceedance

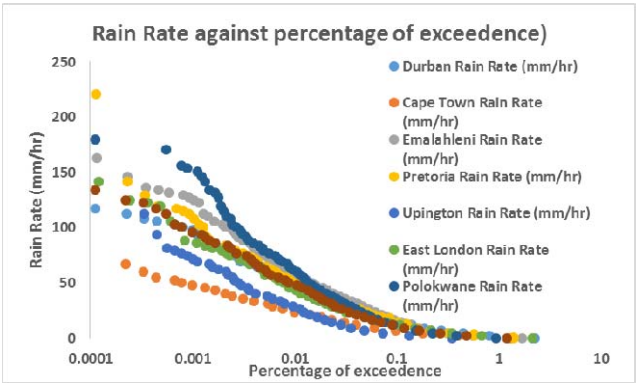


Fig. 3. Rain rate against percentage of exceedance in various station

Figure 3 illustrate the rain rate against the percentage of exceedance. In this figure, the rain rate decreases exponentially as the percentage of exceedence increases. in various stations in South Africa. It is observed that the increase in rain rate, the trend of the sloped curve gradually decreases. The exceedance bends demonstrates that the wet seasons have higher rain exceedance and moderately high rain-rates when contrasted and dry seasons. A noteworthy element of the rain-rate exceedance bend for the locale is the nearness of the breakpoint, which happens at a high rain-rate. The event of breakpoint implies that the rain structure changes bit by bit, with expanding rain-rate, from stratiform to for the most part convective and is more obvious amid the dry season.

Upington, Durban, Polokwane, East London, Bloemfontein, Emalahleni, Pretoria and Cape Town have various rate at 0.01% exceedence are as follows: 29, 53, 77, 43, 48, 60,50and 24 respectively.

Link margin versus link distance

Figure 4 indicate the link margin against link distance with various wavelength of 650nm, 750nm, 850nm, 950nm and 1150nm. It is observed in the figure that as link margin decreases with the link distance increases. At all transmitted power level the scope of connection edge variety is kept up.

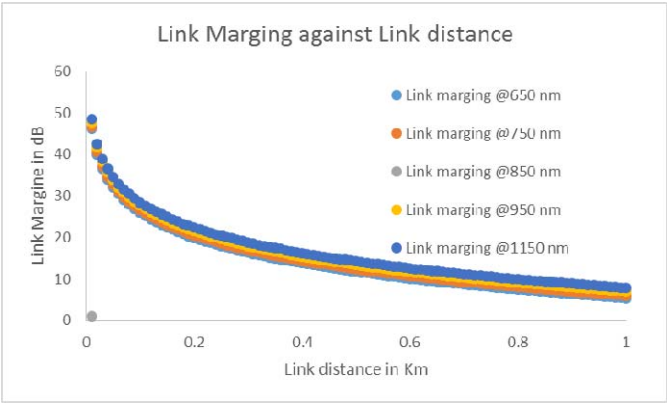


Fig. 4. Link margin against link distance at various wavelength

V. Conclusion

The paper investigated the theoretical analysis of climatic effect (rain, snow and fog) on the FSO communication performance using NRZ-OOK modulation techniques. The link performance is carried out by considering the attenuation contributions by rain, fog and estimate the link margin of the systems against the link distance. The appropriate choice of wavelength has a strong impact on the attenuation coefficient, which leads to elongated transmission in free space. By increasing the transmission, distance will result in decreasing the receiving signal power.

The future work will focus on the impact of the seasonal weather variation on the link performance and address the strength of the weather impairments and proposed mitigating techniques; hence, quality of service will be enhanced.

Acknowledgment

I would like to acknowledge Tshwane University of Technology for the opportunity afforded to me to register for my Doctoral degree and my supervisors Prof P. A Owolawi, Prof S. Ojo and Prof M. I. Mphahlele for their guidance in this research.

References

[1] D. M. Forin, G. Incerti, G.M. Tosi Beleffi, A.L.J. Teixeira, L.N. Costa, P.S. De Brito Andre, 'Free space optical technologies' in Bouras, C.J. (ed.): 'Telecommunication book', IN-TECH, March 2010.

[2] A. A. Basel and R. Islam, "Availability Assessment of Free-Space-Optics Links with Arin Data From Tropical Climates". Journal of Lightwave Technology, vol. 35, No. 19, 2017.

[3] Manor, Haim and Arnon. "Performance of an optical wireless communication system as a function of wavelength." Applied Optics, 42(21), pp. 4285-4294, 2003.

[4] Mohd Supa'at, A.S, Mohammad, A.B, Tong, Y.T and Idrus, S.M. "Unguided optical communication: design and evaluation in Malysian weather." In Proc. Research Seminar on Electronics, Aerospace, Information Technology and Telecommunications, University of Technology Malaysia, pp. 268-271, 1998.

[5] M. Achour, "Simulating atmospheric free space optical propagation: Rainfall attenuation,". High-Power Lasers Appl., vol. 4635, pp. 192-201, 2002.

[6] M. Ijaz, Z. Ghassemlooy, J. Pesek, O. Fiser, H. Le Minh and E. Bentley, "Modeling of fog and smoke attenuation in free space optical communication link under controlled laboratory conditions." J. lightw. Technol., vol. 31, no. 11, pp. 1720-1726, 2013.

[7] A. Basahel, W. Al-Khateeb, M.R. Islam, Al-Khateeb and A. Suriza, "Availability analysis of terrestrial free space optrical link under the impact of rain condition," in Proc. Int. Conf. Comput. Commun. Eng., 2014, pp. 169-172, Kuala Lampur, Malaysia.

[8] A. Vanderka, L. Hajek, J.L. Jan, J. Vitasek and P. Koudelka, "Design Simulation and testing of the OOK-NRZ Modulation format for free space optic communication in a Simulation box", Optics and Optoelectronic, vol. 12, no. 6, pp. 604-616, 2014.

[9] A. Kaur and A. Sheetal, "Performance Analysis of 16 x 2.5Gb/s FSO System for the Most Critical Weather Conditions". Int. Journal of Advanced Research in Computer and Communication Engineering. vol. 4, no. 5, 2015.

[10] S. Rajbhandari, Z. Ghassemlooy, J. Perez, H. Le Minh, G. Kandus3 and V. Kvicera, " On the Study of the FSO Link performance under controlled Turbulance and Fog Atmospheric conditions". 11th International Conference on Telecommunication, ConTEL, Graz, Austria, pp. 223-226, June 2011.

[11] S. Ishii, "Rain Attenuation at Terahertz," Wireless Engineering and Technology, vol. 1, no. 2, pp. 92-95, 2010.

[12] J. S. Marshall and W.M. Palmer. "The distribution of raindrops with size." Journal of Meteorology, 5(4), pp. 165-166, 1948.

[13] W. Zhang and N. Moayeri, "Power-law parameters of rain specific attenuation." IEEE 802.16 Broadband Witreless Access Working Group, Oct. 26, 1999.

[14] A. A. Basahel, and R. Islam, "Availability Assessment of Free-Space-Optics Links With Rain Data From Tropical Climates". IEEE, Journal of Lightwave Technology, vol. 35, no. 19, 2017.

[15] T. H. Carbonneau and D. R. Wisley, "Opportunities and Challenges for optical wireless: the competitive advantage of Free space telecommunications link in today's crowded marketplace." Wireless Technologies and Systems: Millimeter Wave and Optical, Proc. SPIE, vol. 3232, pp. 119-128, 1998.

[16] M. H Mahdieh and M. Pournoury. "Atmospheric turbulence and numerical evaluation of BER in free Space Communication", Optics & Laser Technology, vol. 42, no. 5, pp. 55-60, May 2010.

[17] V. Kvicera, M. Grabner and J. Vasicek, "Assessing Availability performance of Free Space Optical Links from Airport Visibility Data", IEEE phonic Technology Letters, vol. 102, no. 08, pp. 562-565, June 2010.

[18] P. Kaur, V.K. Jain and S. Kaur, "Performance Analysis of FSO array receiver in presence of Atmospgheric Turbulence", IEEE photonic Technology Letters, vol. 124, no. 12, pp. 1165-1168, June 2014.

[19] N. A Mohammed, a.S. El-wakeel and M.H. Aly, "Pointing error in fso link under different weather conditions," International Journal of Video % Image Processing and Network Security IJVIPNS-IJENS, vol. 12, pp. 6-9, 2012 .

A Framework of Bimodal Biometrics for E-assessment Authentication Systems

Temitope Oluwafunmilayo Adetunji, Tranos Zuva and Martin Appiah
Information and Communication Technology
Vaal University of Technology
VanderbijlPark, South Africa
topeadetunji15@gmail.com

***Abstract*—The global use of the internet has improved the growth of the educational sector over the years, while electronic assessments have turn out to be one of the major tools in the development of both non-academic and academic establishments. The effective assessment of a student is mostly perceived as one of the foremost challenges that is frequently experienced during online examination in that it can be very difficult to provide accurate user authentication. The requirement to secure and authenticate a user during e-assessments owing to the high rate of misconduct has led to the proposal of this research. The purpose is to examine potential threats to student authentication during e-assessments and propose a framework which uses a bi-modal authentication approach to provide successful authentication during e-assessment. In implementing this approach, we propose a framework that provides security to improve e-assessments by introducing authentication classifiers to demonstrate its application in biometrics technologies. The proposed model was evaluated based on set of thresholds using Accuracy, FAR and FRR as performance metrics. the proposed model gave a high accuracy of 94.52%. The single-modal model of keystrokes had percentage accuracy of 92.025% and face had percentage accuracy of 92.58%. This implies that the bimodal model integrating keystrokes and face outperforms the single-modal model of keystrokes and single-modal model of face respectively.The study concludes that the proposed model contributes to existing works on e-assessment systems by integrating keystrokes and face bimodal biometric to optimally minimize fraud and impersonation thereby providing accurate authentication of a user.**

Keywords—authentication, biometrics, e-assessments, face recognition, keystroke dynamics

I. INTRODUCTION

E-assessment is regarded as a key activator of student learning. The advantages of E-assessment include: quicker announcements of exam results, increase in test evaluation objectivity [1] , ability to be linked to other computer based or on-line materials and the ability to evaluate courses strengths and weaknesses. These advantages have led many organizations in the education, financial and economic sectors [2] to incorporate E-assessments into their business operations. Thus, in the research of [3], they noted that there are security and privacy concerns with traditional physiological biometric systems as it can be compromised by fake users which pose a threat and risk to E-assessment system. The accustomed ways of distinguishing the identity of an individual by using passwords, pins or cards are not altogether dependable, because they can be lost, forgotten, stolen, revealed, or transferred [4]. However, security and authentication remain a serious concern and various intellectuals have proposed the use of Biometrics to improve security in E-assessment systems. Biometric systems are systems that use computers to identify an individual based on behavioral and physiological characteristics e.g. voice, signature, fingerprint, face etc. [5]. Biometric technology, which is based on physical and behavioral features of human body such as face, fingerprint, hand shapes, iris, palm-print, keystroke, signature and voice, [4, 6, 7] is considered an alternative to existing systems in a great deal of application domains.

The framework presented in this paper uses a bi-modal authentication approach to provide successful authentication and security during e-assessment. The bimodal approach applied in this paper involves two biometric characteristics such as: behavioral biometric characteristics (keystroke dynamics) and a physical biometric characteristic (face recognition).

The keystroke dynamic technique introduced to the e-assessment system refers to the usual arrangements or measures an individual display while inputting on a keyboard input device for example phone and touch screen panels [8], while face biometric recognition is useful for authentication that recognizes face. Face recognition has been widely applied for personal recognition purpose most especially in some areas where high security is of importance. The reasons for its wide applications are due to its intrusive and lower cost compared to other biometric system. In spite of the alternatives of using biometrics technology as an authentication measure, some are still attacked by many challenges that lead to system performance degradation with respect to identification time and accuracy [9]. Though, these problems have been tackled by many researchers using different techniques in order to enhance the overall identification system performance, the ideal solutions for some of these problems are still unavailable. However, in this paper, a framework for bimodal biometrics authentication for e-assessment is recommended and the recommended structure would be of immense importance in that, it will provide successful authentication and security during e-assessment.

II. AUTHENTICATION ALGORITHMS AND CLASSIFIERS

The classifiers that are majorly in use for different biometrics are support vector machines (SVMs) with different kernels (especially Gaussian and polynomials), Gaussian mixture models-based classifiers, neural networks and multilayer perceptron [10-12]. The significant performance improvements of most of them were provided, but the available datasets are what their results depend on. These classifiers play an essential role in providing solutions for problems encountered while using some of the biometric methods. Other techniques proposed and implemented by other researches will be indicated in the literature review.

A. SVM-Based Classifier

SVM-based classifier is used in the most authentication techniques for evaluating the performance of a proposed system.

A SVM is binary classifier that optimally separates the two classes of data. Two major phases require development of SVM as classifier. The first phase involves the determination of the optimal hyperplane which will optimally separate the two classes and the other is transformation of non-problem into linearly separable problem.

1). Applications of SVM to Biometrics

In the work of [13] an iris recognition system was developed, which was confirmed using database of grayscale eye images in order to authenticate a person. In order to develop the user model based on individual's iris code data, the Support Vector Machine was implemented as classifier. Using the Chinese Academy of Sciences–Institute of Automation (CASIA) an experimental study database was carried out to estimate the efficiency of the proposed system. The results obtained indicate SVM classifier produces excellent False Accept Rate (FAR) rate for both open and close set condition. However, future research has to be carried out in order to improve level of usability in order to be able to reduce the rate of False Reject Rate (FRR).

A new framework for Iris and Fingerprint Recognition was proposed by [14] using SVM Classification and Extreme Learning Machine Based on Score Level Fusion. In order to match the features, two biometric features were used so as to develop a bimodal biometric authentication system, these are; iris and fingerprint while, K-mean clustering was used to examine the database. In order to recognize the biometric behaviors, extreme learning techniques and Support vector machine were used in this system. The result obtained indicate a better performance and evaluation of extreme learning techniques and Support vector machine based on score-level fusion methods. Using the two models, no performance analysis of each modality was obtained.

B. Genetics

1) Genetic algorithm based recognition method

What is considered to be part of evolutionary algorithms which is also inspired by Darwinian evolution mechanisms is what can be identified as Genetic algorithms. Genetic algorithms are non-deterministic techniques that carry out modification and crossover operators for developing offspring. The ability to exploit is in the power of Genetics Algorithms, it accomplishes that in highly resourceful manner and it is able to acquire information about an enormous number of individuals [15].

2) Applications of genetics to biometrics

While addressing the problem of low quality fingerprint images, [16] proposed an ideal global transformation with fitness purpose in order to improve the chances of identification between images that possesses two fingerprint.

In the work of [17] they presented a reformed BioGINA algorithm for biometric recognition and the presentation was authenticated with BioGINA database of 1200 fingerprints from [16] and database from 2000, 2002, 2004 fingerprint authentication contests. The outcome of the result indicates an improvement rate of 40% of the equal error rate with fingerprint authentication contest database of 2000.

Additional techniques and applications of biometrics as it applies to authentication will be further explained in subsequent literatures which are presented below:

III. BIMODAL BIOMETRICS

Bi-modal biometrics refers to the use of more than one biometric feature for person recognition. A bi-modal biometric system encompasses the necessary processing required to incorporate the chosen more than one biometric characteristic into the authentication procedure. The use of more than one biometric feature has greatly increased the reliability of the person authentication process. Bi-modal biometric systems help to achieve an increase in performance that may not be possible using a single biometric indicator [18]. It is very difficult for an impostor to concurrently impersonate the various character traits of a rightful user due to the emergence of more than one biometric system [19].

Over the years, in a bid to address the problems of fraud and impersonation in e- assessments, several researchers had worked on more than a single biometric recognition system because of some limitations encountered in the course of their research they believed multi biometric system can address [20].

In view of the above, numerous bimodal approaches with different mode of operations have been proposed in the literature. These include:

The use of palm patterns as well as user name and password in the work of [21] was used. Experimental results of their technique show that both techniques can be used for validation and continuous verification and that the method was good for validation when the learner logs in. Their results further led to the conclusion that the use of a single method of Biometrics is predisposed to imitation and scam [22]. While using timestamp as well as IP address in determining a learner's authentication in an e-exam, [23] designated a new technique which was projected to be used to observe learners. In this experiment, the learner's IP address and timestamps were used to detect any falsified conduct. In their result, they observed that the implementation of the method was stress-free as there is no need for extra hardware but as mentioned in their results, they reached a conclusion that this technique is prone to

imitation threat if the same machine is being used and also, the IP address could easily be changed.

Shende et al. in their research initiated the use of challenge queries by proposing a new method called Profile Based Authentication Framework (PBAF) [24]. The challenge queries were initialized to validate learners that partake in E-assessment. In their results, they observed that using the questions can be easily structured and no extra hardware is needed. However, it was concluded that the method being used is liable to imitation threat because it is considered easy to share. In the work of [25] another technique for authenticating E-assessment was implemented which is voice recognition. In the course of their research, a microphone was used to enhance the effective result on voice recognition to enable the user register for E-assessment. In their result, they discovered that voice recognition can be used for incessant validation and that it can be useful for the blind or people who don't use a keyboard for interaction with the system. However, in their end result, they found out that using a single technique of biometric for E-assessment verification is liable to impersonation and threats.

In the research work of [26], they made use of another technique to detect authentication of users in e-assessment i.e. face recognition. In their research, they tried to examine the precision of face recognition taken on-line using a webcam. In their discovery, they found that face recognition can be used for uninterrupted verification of users. However, they further posited that using one single kind of biometric technique for on-line assessment is liable to imitation and threat [19].

Using mouse movement and finger print in detecting the authentication of on-line users in an on-line examination, [27] in their research work projected a verification system which would be used to sustain several on-line assessment services. In their result, they were able to ascertain that fingerprint technique can be used to detect individuals and their characteristic behavior and can also be used for constant user validation. Their end result however shows that there is a distinct requirement needed for efficient continuous validation and that an essential requirement is a fingerprint enabled scanner device. This devise can pose a challenge to this method of authentication because of its high cost and the fingerprint feature is also liable to theft.

IV. PROPOSED BIMODAL BIOMETRIC SYSTEM

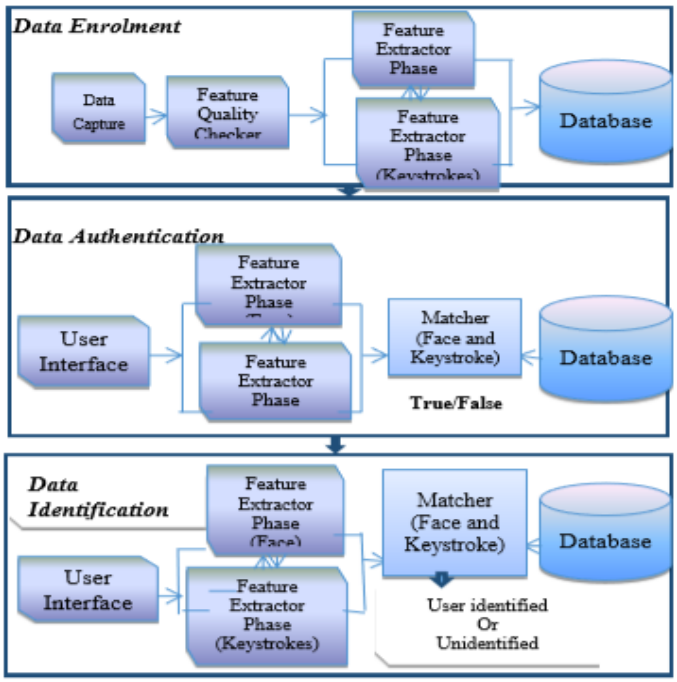


Figure 1: Proposed Framework of Bimodal Biometric for E- assessments Authentication

A. Framework discussion

The following section will discuss the different phases of the framework in figure 1.

1) Data enrolment phase

In the data enrolment phase indicated in figure 1 above, data capture is the first operation to be carried out in E-assessment biometric technique. This stage helps to retrieve data from authorized users using various techniques depending on the biometric features selected. In this research, the biometric feature selected is face because of its advantages of: better security, easy integration, high success rate and automated facial system [28]. It is important because data which is to be authenticated needs to be registered in the database as a template linked internally to a user ID. In the course of the enrolment phase, an individual's biometric features and keystrokes are first perused by a biometric reader, in order to get its digital representation. A quality check is accomplished by the system in order to confirm that continuous stages can consistently practice the developed model. Feature extractors is used to facilitate matching which is then used to develop the sample of data input already in the system in order to create a compressed but sensitive picture known as a template [29]. The biometric system may then pile the template in its central database or record it on an individual's smart card.

2) Data identification phase

In the identification stage, an individual is identified by the system by examining the complete template database for a match, based on facial and keystrokes biometrics. A one-to-many evaluation is shown by the system in order to create the uniqueness of an individual; the process may fail if the individual's ID is not enrolled in the database of the system. This kind of phase is a critical component of both positive and negative identification. The purpose of positive recognition is

for accessibility because the identity of the user may not certainly be demanded; more so, the traditional techniques of personal recognition like PINs, tokens, passwords are activated by positive recognition. Negative recognition on the other hand is initialized to stop a particular individual from using numerous identities and only Biometrics can be used for personal recognition in negative recognition [29].These two operations is to assist the system to limit errors that can pose treat to the E-assessment system.

3) Data authentication phase

Data Authentication is a procedure of verifying the legitimate privileges of an individual before the release of secure resources. This process is achieved by counter-checking data that are distinct which the individual supplies based on the biometric features and keystroke dynamic. The task carried out here on a user's identity needs the system to verify the authenticity of the clone based on his/her biometric features and key stroke dynamics. In this technique, facial biometric technique is proposed alongside the keystroke because of its advantages of; uniqueness, universality, permanence, collectability, acceptability, circumvention and performance [24].

V. PERFORMANCE EVALUATION METRICS

The performance of the system was measured using the accuracy metric as in equation illustrated in equation (1)- (3) below:

$$AC = \left(100 - \frac{\left(FAR(\%) + FRR(\%)\right)}{2}\right) \quad (1)$$

Where $$FAR = \frac{\sum_{i=1} W_{ai}}{T} X100\% \quad (2)$$

$$FRR = \frac{\sum_{i=1} W_{ri}}{T} X100\% \quad (3)$$

W_a is the wrongly accepted individual

W_r is the wrongly rejected individual

T is total no. of users experimented with

VI. EXPERIMENTATION AND RESULTS

The publicly available database of faces was used in this experiment [30]. The database contained 116 faces of 63 men and 53 women. The image resolution used is 768 by 576. Each image was used to query the database but with four varying illumination. Each facial image was mapped to keystroke data therefore only 100 sample members were used since keystroke had only 100 users [31]. The threshold of the system was kept constant. The bimodal system worked in series in that a face was queried then the keystroke data was then requested by the system if the face image was deemed positive. Both biometric systems were supposed to indicate positive for an accurate identification. The table below shows the FAR, FRR and Accuracy of keystroke, face and bimodal biometrics.

TABLE I. RESULTS

S/N	Biometric Model	Average FAR	Average FRR	ACCURACY (%)
1	Keystroke	7.82	8.13	92.025
2	Face	6.927	7.91	92.58
3	**Bimodal**	**1.71**	**2.04**	**94.52**

It can be seen from the table that the proposed model gave a high accuracy of 94.52%. This implies that the bimodal model integrating keystrokes and face outperforms the single-modal model of keystrokes and single-modal model of face respectively.

VII. LIMITATIONS

Both keystroke biometrics and face recognition exhibit some characteristic challenges. Keystroke arrangements are greatly dependent on keyboard layout, device and posture etc. One of the determining factor on the user's typing speed and pressure is the emotional state of the user.[8] found out that when a user is in extremely bad mood, this will result to a reduction in typing speed and pressurewhen compared to an increase in the typing speed and pressure when the user is in good mood. This implies that emotions can have a major effect on typing, as reported by [24]. In the same way, face recognition is extremely sensitive to the conditions of lights, in that, it gives poor results when the light in the environment is in low state. Likewise, over a period of time, physical changes may arise such as distortion of face, wearing spectacles, growing a beard etc. which can result to misclassification of images.

VIII. CONCLUSION

As basis for authentication on any user or student, e-assessments has always generated different forms of inaccuracy, insufficiency and ineffectiveness. This led to invigilators inability to provide accurate user authentication. It is in this regard that this research was conducted. The proposed bimodal biometric model has contributed to existing works on e-assessments by integrating keystrokes and face bimodal biometric to authenticate an individual. The work is unique as it can be used in any academic or non-academic establishments to reduce and minimize impersonation. It is suggested that further research work be directed towards estimating the performance of the proposed framework on a much more enormous database to ascertain the innumerable parameters like equal error rate, complexity of the system, user authentication cost, implementation cost, cost, true negative rate, etc. which are majorly the brain behind a result to approve a validation model.

REFERENCES

[1] Karami, M.e.a., Advantages and Disadvantages of Electronic Assessments in Biomedical Education. IFMBE Proceedings, 2009. 25/XII:: p. p. 61-64.

[2] Alwi, N. and I.S. Fan, Threats analysis for e-learning. International Journal of Technology Enhanced Learning, 2010. 2(4): p. 358-371.

[3] Sharif, M., et al., Sub-Holistic Hidden Markov Model for Face Recognition. Research Journal of Recent Sciences, 2013. 2 (5): p. 10-14.

[4] Zhang, D., Automated biometrics technologies and system. Kluwer Academic Publishers, 2008.

[5] Shende, P.M., D.M.V. Sarode, and P.M.M. Ghonge, A Survey Based on Fingerprint, Face and Iris Biometric Recognition System, Image Quality Assessment and Fake Biometric. 2014(April 2014).

[6] Lin, C. and K.C. Fan, Triangle-based approach to the detection of human face Pattern Recognition. 2011. 34: p. 1271-1284.

[7] Zhang Baochang and et al, Histogram of Gabor Phase Patterns (HGPP). A Novel Object Representation

Approach for Face Recognition. IEEE Transactions on Image Processing,, 2007. vol. 16, No.1: p. pp 57-68.

[8] Yu, Z., D. Y, J. A. K, , Keystrokes Dynamics For User Authentification. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2012(Computer Vision and Pattern Recognition Workshops (CVPRW), 2012).

[9] Awad, A. and A.E.H.e. Al, Machine Learning Techniques for Fingerprint Identification: A Short Review. 2012: p. p. 524–531.

[10] S.Theodoridis and K.Koutroumbas, Pattern Recognition. IEEE Transactions on Pattern Analysis and MachineIntelligence,, 2009. Elsevier,4thedition.

[11] Thom´e, T.C.M.A.C.G., One-against-all-based multiclass svm strategies applied to vehicle plate character recognition. IEEE International Joint Conference on Neural Networks(IJCNN'09), 2009: p. 2153–2159.

[12] Zhao., L., et al., Face recognition based on multi-class SVM. IEEE International ConferenceonControlandDecision, 2009: p. 5871–5873.

[13] Hasimah and Momoh, Iris Recognition using Support Vector Machine. Biometric Systems, Design and Applications. ISBN: 978-953-307-542-6, 2011.

[14] Sangeetha, S. and N. Radha, A New Framework for IRIS and Fingerprint Recognition Using SVM Classification and Extreme Learning Machine Based on Score Level Fusion. Proceedings of 7th International Conference on Intelligent Systems and Control (ISCO 2013), 2013: p. 183-188.

[15] J. Wang, P. Neskovic, and L. Cooper, Training Data Selection for Support Vector Machines. Lecture Notes in Computer Science, Berlin: Springer,, 2005. vol. 3610: p. pp. 554–564.

[16] Tan, X. and B. Bhanu, Robust Fingerprint Identification. in Proc. of IEEE Workshop on Applications of Computer Vision, Orlando, Florida, USA,, 2002: p. pp. 278-280.

[17] Scheidat, T., et al., Optimization for Biometric Fingerprint Recognition Using Genetic Algorithms. in Proc. of the 8th workshop on Multimedia and Security, USA: ACM,, 2006: p. pp. 130–134.

[18] BY. Souheil, Multi-Modal Data Fusion for Person Authentication using SVM. IDIAP Research report, 2007: p. Pgs 1 – 9.

[19] Jain, A.K. and A. Kumar, Biometrics of Next Generation: An Overview to Appear in Second Generation Biometrics. Springer, 2010.

[20] Arun Ross., Salil Prabhakar., and Anil K. Jain, An Introduction to Biometric Recognition. 2004.

[21] Al-Saleem, S.M. and H. Ullah, Security Considerations and Recommendations in Computer Based Testing. Sci. World J., 2014.

[22] Apampa, K.M., et al., User Security Issues in Summative e-Assessment security. 2010. vol 1: p. p. 113.

[23] Gao, Q., Biometric Authentication to Prevent E-cheating. Instructional Technol, 2012. Vol 3.

[24] Ullah, A., et al., Using Challenge Questions for Student Authentication in Online Examination. Int J. Inf, 2012. Vol. 5.

[25] Rudrapal, D., et al., , Voice Recognition and Authentication as a Proficient Biometric Tool and its Application in Online Exam. Int. J. Comput. Appl, 2012. 39: p. p. 6-12.

[26] Penteado, B.E. and A.N. Marana, A Video-Based Biometric Authentication for e-Learning Web Applications. In: Enterprise Information Sytems Filipe, J. and J. Cordeiro (Eds.). Springer, Berlin, Heidelberg, Germany, 2006: p. p. 770-779.

[27] Chellappan, A.a., N.A. Karim, and Z. Shukur, Review of User Authentication Methods in Online Examination. Asian Journal of Information 2008.

[28] Shende, P.M., D.M.V. Sarode, and P.M.M. Ghonge, A Survey Based on Fingerprint, Face and Iris Biometric Recognition System, Image Quality Assessment and Fake Biometric. International Journal of Computer Science Engineering and Technology (IJCSET) / April 2014, 2014. Vol 4(Issue 4): p. 129-132.

[29] Salil, P. and P. Sharath, Biometric Recognition: Security and Privacy Concerns. IEEE Security and Privacy, 2003.

[30] Martinez, A.R. and R. Benavente, The AR face database. TechnicalReport24, ComputerVision Center(CVC) Technical Report, Barcelona, 1998.

[31] http://personal.ie.cuhk.edu.hk/~ccloy/downloads_keystokes100.html.

Business Analytics for Institutional Academic Management: A Case of South African Higher Education

Tshinakaho Relebogile Seaba
Informatics
Tshwane University of Technology
Pretoria, South Africa
seabatr@tut.ac.za

Mmatshuene Anna Segooa
Informatics
Tshwane University of Technology
Pretoria, South Africa
segooama@tut.ac.za

Billy Mathias Kalema
Informatics
Tshwane University of Technology
Pretoria, South Africa
kalemabm@tut.ac.za

Raymond Kekwaletswe
Informatics
Tshwane University of Technology
Pretoria, South Africa
raykekwaletswe@gmail.com

***Abstract*—Data infrastructure and quality are vital organs influencing the health of any organizations and are essential for creating and delivering business insights. Due to this fact, stakeholders expect a flawless experience, real-time solutions and support. However, these expectations are normally too high for IT departments that are always immersed with various users' requests. The Big Data era and its analytics have put organizations in positions to predict and forecast users' needs to produce a great user experience. This is achieved through intuitive design, error-free coding and quality performance. There is increasing demand for better business analytics so as to enable organizations establish solid foundation by building out a data management ecosystem that delivers flexibility and performance required by cognitive solutions. The objective of this paper was to highlight how universities can leverage better analytics to process amount of continuously generated data from various sources and generate actionable insights needed to achieve academic business goals. This study followed an interpretive paradigm taking a case study of a South African university and employed semi-structured interviews. This study found that data value and the use thereof in the management of academic divisions of the university should be consistent with faculty business plans. In this way, business analytics is not haphazardly used but is strategically aligned with faculty plans and overall institutional strategic objectives and targets**

Keywords—Business Analytics, Institutions of Higher Learning, Decision Making, Management

I. INTRODUCTION

Despite the fact that Business Intelligence and Business Analytics complement one another and are, in many instances, used interchangeably, Business Intelligence does not particularly guarantee that you are also going to get Business Analytics. Business Intelligence is a reflective analysis that creates reports based on the current state of affairs and serves as a rear-view mirror that provides business reporting with retrospections of current and previous organizational events. Conversely, while Business Analytics is also concerned with data analysis, it is future-oriented and predictive in nature in the provision of insights on the business and the likelihood of events [1]. Researchers have provided a variety of different definitions for Business Analytics. Davenport and Harris [2] define business analytics by putting emphasis on the collection and application of data analysis technologies and mathematical models that extract every form of value from business processes. Business Analytics is also defined as a process of creating mathematical decision models based on organisational data which leads to data oriented decision making; subsequently ensuring that the use of data adds value to an organization [3]. Business Analytics is furthermore defined in terms of its uses of an array of data types and the fact that its nature can be defined based on business objectives and approaches used in the identification of business value from data [4], [5]. It appears from these definitions that Business Analytics encompasses technologies, models and big data however it is primarily characterized by its ability to focus on the extraction of value from business processes. This means that Business Analytics is not merely about "Big Data" but the use of technological resources with respect of business objectives and value-adding data.

The above description of Business Analytics, specifically [4], [5], creates a link between Big Data and Business Analytics, hence a brief exposition of Big Data Analytics is relevant in this study. [6] note that Big Data (BD) Era has been evident in many communities from government, health to ecommerce organisations which use variety of Big Data to gain detailed insight of their organisation. Big Data is therefore used to leverage opportunity offered by massive of information that comes in Exabyte's and terabytes from social media, sensors and iPhone to gain competitive advantage.

Using data to discover insights for decision making has always existed in organizations to study customer's behaviors through discovering pattern and predicting products demands, however the data that was used had limited insights as only structured data was considered which made it easy to even extract and analyse. Furthermore, currently 80% of the data that the organization generate is unstructured. This data comes in diverse format like pictures, diagram, audio, text which is now complex to process and analyse due to the nature in which it is generated [7]. Many organizations universities inclusive, are facing numerous challenges of analyzing the generated data due to the fact that they lack clear understanding of Big Data analytics (BDA). More still, they noted that the current traditional setting and tools that are being used are inadequate for big data analytics, which pose a disadvantage for universities

from gaining better insights and business value from the Big Data to improve efficiency [8].

Further research on Big Data analytics indicate that many universities and other institutions of learning are still faced with challenges of analytics and that few can draw meaningful insights from the voluminous data they collect on a daily basis [7] This inability to link the business problem to BDA solutions leads to poor decision making that make them loose competitiveness and loss of loyalty from their stakeholders [9], [10], [7].

II. FACTORS INFLUENCING BIG DATA ANALYTICS

In order to produce accurate and actionable data there are technical skills required to handle mining of data and the ability to apply analytical methods to analyse the data. Lack of skills and resources were identified as the main barriers to gaining business value from big data analytics in developing countries (BDA) [11]. A study on the adoption factors of Radio Frequency Identification RFID specific to the manufacturing industry. The study looked at adoption studies that used Technology Organisation and Environment theory and listed the attributes that influenced adoption in all TOE context. These studies had similarities in terms of the attributes discovered for adoption success when using TOE. In all the nine studies made from 1997 to 2004. The attributes that were found common within the technology context are perceived benefits, technology competence. Organisation context had similarities on firm size, financial resources, complexity and compatibility. Environment context had similarities on competitive industry, government promotion, regulatory barriers and partner's usage and partner's readiness [12].

An investigated was conducted on the Determinants of BDA adoption determinants specific to emerging economies in Asia underpinned by Diffusion of Innovation (DOI) and Technology-Organizational-Environment (TOE). A quantitative approach was used to validate the developed model. While the size of the organization, competition, environmental instability and compatibility were identified as supporting factors; complexity and regulatory support were identified as the major barriers to BDA adoption [13] – [15].

III. BUSINESS ANALYTICS IN INSTITUTIONAL ACADEMIC MANAGEMENT

With relevance to the purpose of this paper, Business Analytics can be applied to various settings within higher education such as donor tracking, admission processing, recruitment, administrative and instructional applications, financial planning and student performance monitoring [16]. This shows that Business Analytics is not student data centered but spans different divisions of an institutions of higher learning. Thus, data has the capability of creating value in business processes and operations beyond teaching and learning. Business Analytics systems have the capability of optimizing organisational performance and are regarded as major strategic investments.

A study was conducted on decision support systems and how they influence the development of the university's strategic management at the Islamic university in Gaza. A descriptive quantitative research approach and stratified random sample were used in the study, which found that there exists a correlation between the university's strategic management and decision support systems. Their study further highlighted the importance for universities to consider decision support systems. The use of Business Analytics at different management levels leads to a focus on the realization of business goals through reporting of data for trend analysis, business process optimization for enhanced performance and predictive modelling for forecasting [17]. Therefore Business Analytics in the education sector is not merely applied but should be linked to what the institution is all about including goals, mission and vision.

IV. LEVELS OF BUSINESS ANALYTICS FOR INSTITUTION OF HIGHER LEARNING

The levels of Analytics are employed to address different sets of questions to discover insights in order to proactively respond to opportunities and issues in real time.

A. Descriptive Analytics

Descriptive analytics simply shows what happened, through generation of metrics using data for monitoring day-to-day processes or operations [18] and business reporting [19]. This level of analytics answer's to the question that aims to understand what has happened based on the historical data. Within the higher education context, the descriptive analytics can be used to draw reports that shows the pattern in which the university has being performing in the academic perspective. It answers questions such as "For the past 3 years how many graduates the university produced?" [20]; [21].

B. Diagnostic Analytics

Diagnostic Analytics follows descriptive analytics and examines why events in data happened, thus discovering or identifying the root cause of a problem This process in turn yields surprising discoveries [18].This level of analytics therefore responds to the question that seeks to understand the root course of the result received from the descriptive analytics. The question that calls for these levels will be, why did it happen? This level can be used to understand why the institutions obtained low graduate rate for the past 3 years, this insights will help university management to address the issues to improve the graduation rate going forward [20]; [21].

C. Predictive Analytics

Predictive Analytics is a data mining category that combines statistical analytical techniques and business knowledge for the exploitation of hidden predictive information and the prediction of future patterns and trends derived from the analysis of historical data with diverse variables. The variables are the essential elements of Predictive Analytics referred to as predictors, which are then combined into a predictive model [22]. This level suggest the forecast of what could happen in the future based on the trends and pattern in which the organisation has been progressing or performing, the results of the previous years can give a directive if the institutions is highly expected to progress or to regress. Predictive level also relies on the historical data to project the future position of the organisation [20],[21].

D. Prescriptive Analytics

Prescriptive Analytics is decision oriented as it is concerned with the suggestion of actions to be taken for the optimization of business operations and subsequently maintain its position in the industry [18].This level Prescriptive analytics is used to respond to the question, "what is the best suitable solution given the circumstances?" and furfure recommend decisions to make or actions to take as a way of to resolving the issues or to achieve certain goals. [20],[21].

V. RESEARCH METHODOLOGY

A. Research Philosophy

The research philosophy reflects a researcher's assumptions of what "valid research" is. It is from these assumptions that the research strategy and unit of analysis are clarified [23]. This study followed the interpretive ontological assumption as the researchers believed that reality construction is based on perception and actions of social actors [24]. Thus, it was believed that subjective interpretation of the participants involved provided the best way for assessing the role of Business analytics in institutional academic management. While descriptive studies answer questions of "what, where, how", explanatory studies answer the "why" question. This study is descriptive in nature as it was aimed at describing how universities can leverage better analytics to process amount of continuously generated data from various sources and generate actionable insights needed to achieve academic business goals.

B. Research Approach

This study is based on the view of Information Systems discipline as being comprehensively dependent on the application of methods used to answer research objectives [25]. With relevance to this study, the main objective investigated entailed how universities can leverage better analytics to process amount of continuously generated data from various sources and generate actionable insights needed to achieve academic business goals. This includes operationalizing constructs of interest through the selection of appropriate research method, devising an appropriate sampling strategy [26] and analysis of data [27].

For this study, qualitative research approach was adopted due to its strength to systematically and subjectively describe life experiences such as Business Analytics. Business Analytics process was not easy to articulate in an umbrella form, an understanding and description of this process required the employment of face-to-face interactions with executive managers to have an in-depth understanding of the different levels of analytics. The interpretivism paradigm goes hand in hand with the qualitative nature of this study because Business Analytics, as the processes under study, does not have objective meanings and is specific to the environment in which it exists.

C. Research Strategy

This study took a case study of a South African university, following Myer [28]'s case stages, the case study strategy made it possible for this study to determine the status quo of Business Analytics. Additionally, data about the Business Analytics background were collected through semi-structured interviews and finally present an analysis of findings that show how universities can leverage better analytics to process amount of continuously generated data from various sources and generate actionable insights needed to achieve academic business goals.

D. Sampling and Data Collection

Based on purposive sampling technique, academic executive managers from five faculties within a South African university participated in this study. First, the institutional website was used to have an overview of the institution, its background and structures in place. This was followed by drafting of a list of possible participants and contacting relevant secretaries to check availability and request a meeting with executive managers. Once respondents agreed to participate in the study, a date for the interview or meeting was confirmed. An interview pack was sent to the participants' who consented for the interview, in total 13 semi-structured interviewed.

VI. RESULTS AND DISCUSSION

The purpose of the research was to describe the role of Business Analytics for Institutional Academic Management in a South African higher learning institution. In order to achieve the aim of the research, data collection was conducted by interviewing academic executive managers in academic divisions of the university. Thematic analysis was used for data analysis and the findings have been structured in such a way that the role of data analytics is described based on its relevance to teaching, learning and research.

A. Descriptive Data Analytics Reporting

A participant provided an overview of general reports requested by Deputy Vice Chancellor and Vice Chancellor from his portfolio. The overview was provided thus: ["There are many reports such as the success rate reports, because we have subcommittee of success rate and other reports that are needed from me are the earmarked grants report. Each year we have to supply the admin how we have performed with those year mark that they have given and also professional boards. You know some of our programs …….are managed by professional boards so we need to response to their professional boards."]. The interviews reflected descriptive Business Analytics reporting for success rates, earmarked grants and reports submitted to professional boards. In support of these findings, The use of Business Analytics is influenced by its ability to create reports that are accessible and understandable by all users, irrespective of the level of management and operation [29].

B. Data Analytics Flexibility and Agility

One of the academic executive managers elaborated on descriptive reports requested from his portfolio. He had this to say about the data sources from which these reports are generated: ["It depends on the information that you need, for example, they want information of average number of students that have enrolled. Obviously its ITS or either they use the normal Management Information Systems (MIS) that they used before but I think it's ITS data if it's about the students, if it is about finance obviously I will go to ITS (Integrated Tertiary Software) and check."]. Among the vast categories of Business Analytics, real-time Business Analytics creates a dynamic environment, which requires

readiness for a change to prompt decision-making. This is not only time-oriented and emphasized at top management but it is also involves the need for a Business Analytics culture at the operational level. The use of business analytics for different purposes within the faculty indicates the role it plays in instilling an analytics culture and flexibility in decision making pertaining to both academic and administrative activities.

C. *Descriptive Analytics Visualization Technology*

Evident in the following extract: ["There are quite numerous amounts of reports that we get from the (MIS) and so much that what we have done now as an institution we have got a dashboard now called Higher Education Data Analyser (HEDA). Any form of data that you require, I can get that for you as long as I'm in the office and the only problem is that you can't access HEDA from home or outside the university"], majority of the respondents mentioned institutional management information and analytics systems as the primary technologies for the generation of reports. Researchers such as Ramamurthy, Sen, and Sinha [29] supported the use of reporting technologies include Online Analytical Processing (OLAP), dashboards, and scorecards. From the participant's response, it can be argued that there is a need for ease of access to these systems outside the university's physical location.

D. *Diagnostic Analytics for Teaching and Learning innovations and service delivery*

Executive managers in universities do not only focus on research, teaching and learning, but are responsible for ensuring equipment and advancements of academic facilities as part of student service delivery. This finding was extracted from the following excerpt pertaining to the use of data in the improvement of business processes with the faculty: ["I need to go to ITS and check for example, how many students enrolled in this particular classroom, ITS will tell me what kind of Wi-Fi should I install in this classroom depending on number of students that are in the class, so I need data."]. In this respect, the Business Analytics "know how" is of great importance especially for people involved in decision making that relate to innovations and well-improved services and processes that make an organization to stand out in the competitive environment in which it exists [30]. Based on the above response, the role of Business Analytics is evident in its capaility to provide reports necessary for improved delivery of service to students through the improvement of Information and Communication Technology (ICT) infrastructure to accommodate the blended learning approach.

E. *Diagnostic business analytics for performance measurement*

An executive dean was requested to elaborate on the use of institutional analytics systems for reports in line with faculty performance measurement. The responses below indicate the role that business analytics plays in measuring academic performance:

["To be honest with you, I think what we have done and it's a general practice in the university, we said we have got two big targets and that target is success rates for undergraduates and postgraduates. Now, this to the last scale was used as a measurement of performance for success rate, so success rate initially was below 72%...] and [For instance if I'm now ready to find out subject with a success rate of less than 40% and all subject within the faculty I can be able to draw that list and I can also draw a list of student who aren't performing well…"]

Previous research such as [31] indicated that analytics goes beyond merely focusing on generating insightful information from data sources; it goes the extra step to create an improvement in measurable business performance. Faculty-wide success rates in relation to this study are the major performance measurement indicators.

F. *Diagnostic business analytics for at-risk subjects*

An academic executive manager was requested to provide an overview of strategy development within the faculty; as well as the role that data plays in the development of faculty business plans. Based on the participant's response, the role of diagnostic analytics is evident in the use of data to further uncover the causes of academic poor performance, as the following extract shows: ["Then as a faculty we discover that there is/are subject(s) that is offered as a problematic subject or service subject, then we realise that this is the subject(s) that is creating the problem. We then see that the percentage pass of students is low, then the HOD of that department will provide report on what causes this"]. In support of this finding, research such as [32] indicated that diagnostics Analytics in most cases discovers cause of identified problems, leading to further investigations into the business problem using mathematical techniques such as correlations. In this case, it is clear that the faculty makes use of data to justify not only to identify subjects with low pass rates and provide relevant data driven diagnosis. However at-risk students in problematic subjects can subsequently be identified.

G. *Predictive Analytics for risk and cost reduction*

When asked about data use in the identification of academic, the dean in one of the faculties responded thus: ["….the success rate of students is a risk, meaning your teaching input unit, teaching output unit and FTE unit (full time equivalent) needs to be checked because remember if we don't reach our target we are going to be penalized, so I need to check, it's a risk which I have identified and I need to work around it."]

While the above results indicated the use of success rates for diagnostic analytic in terms of academic performance measurement, the same indicator is used for predictive purposes in relation to teaching input unit, teaching output unit and FTE unit (full time equivalent); this is to avoid government penalties for not achieving set targets. Parks and Thambusamy [33] reviewed 2012 Harvard Business Review Analytics which emphasized the role of analytics in reducing risks and costs. These findings show that predictive analytics should be aligned with the achievement of strategic goals and relevant targets.

H. *Prescriptive Analytics and Executive leadership support*

Business Analytics should not be utilized from a single-angle where it is only perceived as a reporting technology.

Instead, all divisions or subunits in an organization should be encouraged to use Business Analytics continuously and incrementally [34]. This encouragement, according to Laursen and Thorlund [35] leads to executives' organization-wide persuasion to embrace and enjoy using Business Analytics. In relation to this study, improvement of student performance is guided by the availability of funds, extracted from institutional data. This finding was extracted from the following response from an assistant dean with pertinence to the use of data during strategy development" ["We use this grant to appoint tutors, then these tutors works with success rate coordinators within the faculty in the office of the dean, working together with the assistant dean for teaching and learning."]. The interview further revealed that through the use of institutional data, the faculty is able to plan ahead after the diagnosis of poor academic performance. Based on the development grant from the Department of Higher Education and Training (DHET), human resource allocation is done to improve success line in line with the performance of the faculty.

These funds are utilized to provide support for at-risk students with the intent to increase faculty success rate. However, from the above response, the involvement of academic executives in using data to support student support initiatives indicates the role of business analytics in academic process improvement as well as executive.

VII. CONCLUSION

Business Analytics poses many opportunities for many organizations, universities included, to gain more insights to better their decisions and improve competitive advantage. This study showed that the business analytics in institutional academic management is primarily based on success rates, grants, teaching input and output units, yet the analytics process depends on availability of analytics technologies. Success rate has proven to be the - indicator in the academic division of the university as it brings out diagnostic analytics for both the identification of at-risk students within the faculty and subsequently indicate the performance of faculties. Success rates have also proven to be the basis of predictive analytics in terms of financial planning and cost reduction. As much as reports are required by academic executive managers, the dynamism characteristic of South African higher education requires a business analytics culture that enables managers to make quick decisions from the strategic, business and operational levels of the academic divisions. This creates a culture in which different kinds of reports are used for academic and operational decisions in line with higher education challenges and the turbulent environment in which South African universities operate.

While South African universities are expected to be innovative, at the same time provide improved teaching and learning services to students, analytics should not be overlooked but could be used as a driver towards innovations. Moreover, the findings show that prescriptive analytics plays a role in improvement of academic performance. This study showed that different levels of analytics are applicable for different focus areas in the academic division of the university. The results also show that institutions of higher learning are already taking advantage of business analytics to not only draw intelligence but also mirror the university performance and further interrogate the data to understand why the performance is at a particular standing point. In the management of academic divisions, the value of data and the use thereof should be consistent with faculty business plans. In this way, business analytics will not be used haphazardly but will be strategically aligned with faculty plans and overall strategic objectives and targets of the university.

RECOMMENDATIONS

It could be more beneficial for higher institutions of learning to utilize even the silent data that is generated in the background when academic activities are performed. This data can give more insights around the behavior, engagements and students sentiments which can uncover the likelihood of the student success or failure. Such silent data can be drawn from different sources, generated in high velocity and volume. Therefore, further research should be conducted to fill this gap.

REFERENCES

[1] F. Kalota, "Applications of Big Data in Education," *International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering*, vol 9 (5), pp. 1602–1607. 2015.

[2] T. H. Davenport and J. G. Harris, *Competing On Analytics: The New Science of Winning*. Harvard Business School Press, Boston, Ma, 2007.

[3] J.O. Scherer, A.P. Kloeckner, J. L. D. Ribeiro, G. Pezzotta, and Pirola F. "Product service System (PSS) Design: Using Design Thinking and Business Analytics to Improve PSS Design," *Procedia* CIRP, vol 47, pp. 341-346. 2016.

[4] H. Demirkan and D. Delen, "Leveraging the Capabilities of Service-Oriented Decision Support Systems: Putting Analytics and Big Data in Cloud, " *Decision Support Systems*, vol 55 (1), pp. 412–421. 2012.

[5] C. Holsapple, A. Lee-Post, and R. Patkath, "A Unified Foundation for Business Analytics, " *Decision Support Systems*,vol 64, pp. 130-141.2014.

[6] H. Chen, R.H. Chiang, and V.C. Storey, "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS Quarterly*, vol 36(4), pp. 1165-1188. 2012.

[7] B. Daniel, "Big Data and Analytics in Higher Education: Opportunities and Challenges, " *British Journal of Educational Technology*, vol 46(5), pp. 904-920. 2015.

[8] H. Brown-Liburd, H. Issa, and D. Lombardi, "Behavioral Implications of Big Data's Impact On Audit Judgment and Decision Making and Future Research Direction, " *Accounting Horizons*, vol 29(2), pp 451-468. 2015.

[9] D. Agrawal, P. Bernstein, E. Bertino, S. Davidson, U. Dayal, M. Franklin, J. Gehrke, L. Haas, A. Halevy, and J. Han, "Challenges and Opportunities with Big Data 2011-1," Cyber Center Technical Reports, Purdue University, Jan. 1. 2011.

[10] M. Bala and D. Ojha, D, "Study of Applications of Data Mining Techniques in Education," *International Journal of Research in Science and Technology*, vol 1(4), pp. 1-10. 2012.

[11] E.M. Micheni, "Diffusion of Big Data and Analytics in Developing Countries," *The International Journal of Engineering and Science (Ijes),* vol 4(8). pp 44-50. 2015.

[12] Y. M. Wang, Y.S. Wang, and Y.F. Yang, "Understanding the Determinants of RFID Adoption in The Manufacturing Industry," *Technological Forecasting and Social Change*, 77(5):803-815. 2010.

[13] K. Agrawal, "Investigating the Determinants of Big Data Analytics (Bda) Adoption in Asian Emerging Economies," *Academy of Management*, Vol5(1). 2015.

[14] E.M. Rogers, *The Diffusion Of Innovation*, *5th Edition*. New York: Free Press. 2003.

[15] R. Depietro, E. Wiarda, and M. Fleischer, "The Context for Change: Organization, Technology And Environment," *The Processes of Technological Innovation*, vol 199(0), pp. 151-175. 1990.

[16] A.G. Picciano, "The Evolution of Big Data and Learning Analytics in American Higher Education," Journal of Asynchronous Learning Network, vol 6. 2012.

[17] M.J. Shobaki and S. S. A. Naser, S, "Decision Support Systems and Its Role in Developing the Universities Strategic Management: Islamic University in Gaza as a Case Study," *International Journal of Advanced Research and Development*, vol 1(10), pp. 33-47. 2016.

[18] A. Banerjee, T. Bandyopadhyay, and P. Acharya, "Data Analytics: Hyped Up Aspirations Or True Potential?," *Vikalpa*, vol 38(4). 2013.

[19] P.M. Hartmann, M. Zaki, M, N. Feldmann, and A. Neely. "Big Data for Big Business? A Taxonomy of Data-Driven Business Models Used By Start-Up Firms". Working Paper. University If Cambridge, Cambridge Service Alliance. 2014.

[20] M. Feki, I. Boughzala, and S.F. Wamba, "Big Data Analytics-Enabled Supply Chain Transformation: A Literature," *2016 49th Hawaii International Conference on System Sciences (HICSS), Koloa, HI, USA*, January 5-8, 2016, ISSN: 1530-1605, ISBN: 978-0-7695-5670-3, 2016. pp: 1123-1132.2016.

[21] Y. Perwej, "An Experiential Study Of The Big Data," *Science and Education*, vol 4(1), pp. 14-25. 2017.

[22] O. Z. Hashmi and S. Scheikh, "Impact Of Social Attributes On Predictive Analytics In Telecommunication Industry". IEEE. 978-1-4673-2252-2/12. 2012.

[23] M. Myers and D. Avison, D. (2002). *Qualitative Research in Information Systems*. A Reader Sage Publications Ltd. 2002.

[24] M. Saunders, P. Lewis, and A.Thornhill. *Research Methods for Business Students, 4th Ed*, Harlow: Prentice-Hall. 2007.

[25] A. Pinsonneault and K. Kraemer, "The Impact of Information Technology on Middle Managers," *MIS Quarterly*, vol 17(3), pp. 271-292.1993.

[26] A. Bhattacherjee, *Social Science Research: Principles, Methods and Practices.* Textbooks Collection. 3. 2012.

[27] G. Gable, 1994. "Integrating Case Study and Survey Research Methods. An Example in Information Systems," *European Journal of Information Systems*, vol 3, pp. 112±126.

[28] M.D. Myers, "Qualitative Research in Information Systems," *Management Information Systems Quarterly*, vol 21(2), pp. 241-242. 1997.

[29] K.A. Ramamurthy, A. Sen, and A. Sinha, "Data Warehousing Infusion And Organizational Effectiveness, ", *IEEE. Transactions on Systems, Man, and Cybernetics—Part A—Systems and Humans,* vol (38)4, Pp. 976–994. 2008.

[30] R. Sharma, P. Reynolds, R. Scheepers, P. Seddon, P, and G. Shanks, "Business Analytics And Competitive Advantage: A Review and Research Agenda," *In A. Respicio, F. Adam and G. Phillips-Wren (Eds.) Bridging the Socio-Technical Gap in DSS - Challenges for the Next Decade*. Amsterdam: IOS Press. 2010.

[31] M.J. Schniederjans, D. G. Schniederjans, and C.M. Starkey, 2014). *Business Analytics Principles, Concepts, and Applications, What, Why, And How*. Pearson Education, Inc. Upper Saddle River, New Jersey 07458. Isbn-10: 0-13-355218-7 Isbn-13: 978-0-13-355218-8. 2014.

[32] K. Blasiak, K. (2014). "Big Data; A Management Revolution: The Emerging Role of Big Data In Businesses," PHD Theses, Retrieved From: [Https://Www.Theseus.Fi/Bitstream/Handle/10024/74701/Big%20data%20%20a%20management%20revolution_Kevin%20blasiak.Pdf?Sequence=1]

[33] R.F. Parks. R.F. and R. Thambusamy, "Understanding Business Analytics Success and Impact: A Qualitative Study," *Information Systems Education Journal (ISEDJ)*. vol 15 (6). 2017.

[34] T. H. Davenport, J.G. Harris, and R. Morison. *Analytics At Work, Smarter Decisions Better Results*. Harvard Business School Publishing Corporation, Cambridge, Ma. 2010.

[35] Laursen, G.H.N., and Thorlund, J. (2010). Business Analytics for Managers: Take Business Intelligence beyond Reporting. Hoboken, Nj: John Wiley And Sons.

Overview on Intrusion Detection Schemes for Internet of Things (IoT)

Saher Ghayyad and Shengzhi Du
Department of Electrical Engineering
Tshwane University of Technology
Pretoria, South Africa
saher.ghayyad@secureprotocol.co.za

***Abstract*—In our modern world, almost every aspect of our lives is directly affected by the revolution of digitalization. Everything we rely on nowadays has a computer chip and a software controlling it whether it is a mobile phone, a car, a computer, or any other electronic device. These individual life affecting technologies are composing of cyber physical systems communicating via the Internet of Things (IoT). This paper addresses the development of Intrusion Detection Schemes for IoT. The increasing risks in IoT infrastructure compromise mainly against IPv6 over Low Power Wireless Personal Area Networks (6LowPAN) are a concern. The data breaches and data manipulation are discussed by identifying the most recent hacker methodologies and tools and analyzing what the breach will affect inside an IoT network. After discussing the "Problems", the paper will discuss the existing intrusion detection schemes with their limitations and will propose a more effective solution.**

Keywords—Intrusion Detection Schemes, Penetration Testing, 6LowPAN, Defense Strategy

I. INTRODUCTION

For years, the internet has been a major tool in our lives. Almost all electronics equipment we rely on uses the internet to communicate and transmit data from one place to another. This communication mechanism known as the internet has been always a target for hackers who intend to take advantage of individuals, corporates, governments, banks or other institutions.

A new internet mechanism believed to be the next generation of internet has evolved under the name Internet of Things (IoT) where each physical system on this wide network can communicate without any human intervention. These are smart systems known as cyber physical systems that use the IoT network to communicate with each other [1]. A cyber physical system can vary from a smart grid in a power station all the way to a smart phone used by individuals. These systems are attractive to hackers that intend to cause as much damage as possible. This fact imposes high risks and makes the internet of things face lots of obstacles that can lead to a compromise of the cyber physical systems endangering the lives of thousands of individuals worldwide. These risks involve mainly core infrastructure and network communication security. This section will provide a background on the IoT core infrastructure based on the routing protocol for low-power and lossy networks or RPL. This clarifies how the core communication mechanism operates, where the vulnerabilities are and how to protect it from intrusion activities afterwards.

RPL or the IPv6 routing protocol for low-power and lossy networks is the main protocol at the core infrastructure of the internet of things. It is the routing protocol that dominates the communication routing of the IoT small and low power devices throughout the IoT network. This protocol was chosen to be a standard by IETF in 2011 allowing IoT devices to be manufactured and rolled out the market. The IoT small devices can be referred to as nodes. RPL routing topology identifies a main root which will operate as the central node collecting and passing data to the system. During this communication process, all nodes will be operating at the perimeter of a network associated to several parental nodes that are on standby in case the primary main node failed to process data. This process is called a Multipoint-to-Point communication where the data travels from the nodes to the main root node. The opposite process is called a Point-to-Multipoint where the data travels from the main root node to the rest of the nodes on the network. The third traffic pattern will be Point-to-Point between a device and another device. All three mentioned patterns are the patterns supported by RPL. An example of these nodes are sensors that sense the environment around them then communicate the data sensed to the rest of the network [4].

Wireless Industrial Sensor Networks or WISNs are widely used in industrial environment such as chemical and nuclear plants. It is very expensive for these environments to have wired LAN only due to the difficulties because of their harsh conditions. Not all plants are in an ideal location with ideal conditions. This posses challenges for LAN deployments and forces these structures to use wireless communication such as wireless industrial sensor networks. The responsibility these sensors have is immense and any failure can lead to a catastrophe such as a gas leak or even radiation leak. This makes it highly critical for these sensors to be protected against any malicious activity. Moreover, these sensors must have the ability to repair themselves quickly and recover from any fault as fast as possible to minimize the impact or even eliminate it. It has been mentioned in one of the researches done in 2010 under the title "An Experimental Study of Hierarchical Intrusion Detection for Wireless Industrial Sensor Networks" [8], that companies such as Intel have developed a preventive maintenance application which automatically monitors the health of these sensors. They also mentioned that WISNs are used widely in various critical sectors such as the oil sector

and they named British Petroleum as one of these companies which used WISNs to monitor remotely its customer's LPG tanks [8].

This paper will focus on developing an Intrusion Detection Scheme for IoT. It will discuss some of the "problems" of increasing risk in IoT infrastructure compromise mainly against 6LowPAN leading to data breaches and data manipulation by identifying the most recent methodologies and tools hackers uses nowadays and what will happen inside an IoT network once the breach takes place. After discussing these "problems", the paper will focus on the "Solutions" by listing existing intrusion detection schemes available for IoT with their limitations, then develop a more effective Intrusion Detection Scheme by presenting an active in-depth-defense strategy and algorithm against these attacks relying on the intensive research that will be done in this field with the focus on 6LowPAN using penetration testing capabilities. The proposed "Solution" will add value and be part of the most recent studies to combat cyber threats affecting the IoT infrastructure.

The paper is structured as follows: Section I main goal is to introduce the internet existing communication mechanisms as well as the internet of things next generation communication mechanisms used by cyber physical systems based on the Routing Protocol for low-power and lossy networks RPL. Section II will represent the literature review of this paper. This section will provide a background about existing cyber security threats affecting the communication sector mainly the IoT core based on RPL. White, grey and black hackers. Section III will discuss existing hacking techniques as well as tools used to compromise the IoT communications networks. Moreover, based on the footprint activities such as passive reconnaissance, active reconnaissance, vulnerability assessment, exploitation, breach and destruction, the behavior of potential hackers are presented. Section IV will reveal existing intrusion detection schemes and its limitations in combating cyber threats. Section V will reveal a more effective intrusion detection scheme that will add value in combating and defeating these attacks to protect the IoT Infrastructure. It will be considered to model the behavior of potential hackers. Then strategies will be provided to avoid the followed attacks. The machine learning techniques will be employed to classify the suspect behavior and the ones of normal network users. Section VI concludes the paper by the summary of contributions, challenges, and future works.

II. IoT Security Threats

The Internet of Things or IoT is the new generation of internet which is expected to transform the way we live and work. It is structured in a way where all nodes that are interconnected will send and receive data across the IoT communication infrastructure using protocols such as RPL routing protocol as discussed in section I. This communication process must guarantee service availability and data privacy.

The IoT architecture in this paper focuses on the core of four layers;

- ***Sensing Layer*** which sits at the bottom of the IoT architecture where its devices sense the environment and based on this sensing capabilities, it will determine what information to be sent and how.
- ***Network Layer*** which forms the communication structure for all wired or wireless networks to communicate.
- ***Service Layer*** which handles information related to services or information requested by the user to be executed.
- ***Interface Layer*** which processes all interactions between the user and the application [2].

Starting with the sensing layer, where communication happens between nodes, it is critical to have security measures defending against any breach. "At sensing-layer, the devices are designed for low power consumption with constraints resources, which often have limited connectivity" [2] Since at this layer the devices communicate directly with each other, authentication between these devices is crucial. The sensor at this layer needs to be able to sense the right environment and send data accordingly. RFID tags can be used as an example where sensors need to be able to identify the correct person then authenticate and allow access or block access if the user information was not verified. In most cases, these machines are automatically configured to download patches to update the system. The challenge will be in how the sensors are sensing environment's updates and how are they authenticating these patches before being automatically downloaded onto these systems. Threats are always there putting this layer in jeopardy. Some of these threats can be un-authorized access, unavailability, denial of service attack, malicious code and other.

Second comes the network layer of the IoT infrastructure. This is the layer that connects all the devices and allow them to talk to each other. It sits between the sensing layer and the service layer. It can receive data from one layer, combining several elements, then passing it to the next layer. IoT connects networks with each other making it complex when it comes to security. The bigger the network spread, the more vulnerable it becomes. As discussed in section I, RPL is the routing protocol used by the communication protocol in low power and lossy networks called 6LowPAN (IPv6) over Low Power Wireless Personal Area Networks. This sheds the light on a major topic which will be discussed in this section related to the security of the IoT routing infrastructure.

RPL, like any other protocol, has been investigated for vulnerabilities and possible breach scenarios. A study was done by group of researchers around the RPL Ranked Property which provides routing path optimization where it validates the path data must follow during the data flow reducing congestion and control overhead. [4] The study drafted an attack graph which explains the several ways an attacker uses to exploit RPL routing which means exploiting the IoT infrastructure through exploiting the vulnerabilities within RPL Ranked Property. The node design and the type of connectivity within the IoT network in relation to RPL

was discussed in section I. This section discusses the types of control messages RPL uses such as:

- DIO (Destination Oriented Directed Acyclic Graphs Information Object).
- DIS (Destination Oriented Directed Acyclic Graphs Information Solicitation).
- DAO (Destination Oriented Directed Acyclic Graphs Advertisement Objects).

The DIO mission is to send information to the sensor node giving it the ability to discover RPL, identify and learn the Destination Oriented Directed Acyclic Graph DODAG to select the parent node. The DAO mission is to be advertised by a node that it exists by letting the sink node know the destination point to where the traffic must be delivered. The location of a node in this paper is referred to as the rank of a node. In networking topology, we have parent nodes and child nodes. The rank of the child node will always be bigger than the rank of the parent node as it is further out of the core network than the parent node is. The root node usually has the least rank as it usually sits at the core infrastructure of a network while perimeter nodes have the highest rank as they sit at the edge of a network. When the communication process initiates, the child node searches for the closest node to it (the lowest rank) and selects it to be its parent node. This routing topology of how the nodes communicates provides efficient connectivity and optimal network performance.

Internal networks are highly vulnerable to all types of attacks that can be unleashed by a malicious node. In the case of a breach, a malicious node will falsify almost every information transmitted starting with its rank. The malicious node will falsify its rank by decreasing its value fooling the other nodes that will prefer to connect to it assuming it is the closest. This type of attack is called the Decrease Rank Attack. On the other hand, the malicious node can increase its rank fooling a node at the perimeter of the network who assumes a change of parental state where it will try and connect to the malicious node since it was fooled the malicious node is closer to it than the actual parental node. By doing this, the node trying to connect will broadcast a flood of traffic which will be travelling throughout the network causing congestion and consuming extra power etc. This type of attack is referred to as the Increase Rank Attack. There are several types of different attacks which will be discussed in section III of this paper.

More research took place in relation to the 6LowPAN and its Routing Protocol RPL vulnerabilities. A study was done focusing on Denial-of-Service detection in 6LowPAN based internet of things. Cisco predicts that by the year 2020, there will be more than 50 billion devices connected to the internet creating large volumes of data which can reach tera-bytes per second. The large number of IoT devices on the internet utilizing the 6LowPAN and its RPL routing protocol leaves them with limited resources such as low power, weak processing capability, low memory and even unreliable communication due to congestions. This pauses major threats as these weaknesses can be exploited easily by hackers [6]. A good example of easy disruptive attack can be Denial-of-Service attack which will be discussed and demonstrated in detail in section III.

The focus in this paper will be on Wireless Sensors Network in IoT. The regular network infrastructure which uses IP has several security models deployed. An access point can be used as an example where encryption is deployed in the form or security keys such as WPA. These security encryption measures can't be used in the case of 6LowPAN when it comes to IoT Wireless Sensors as the limited power characteristic of these devices makes them unable to handle such encryption models. This leaves the IoT Low Power devices more vulnerable than their counterparts in the IP world. The research processes are ongoing to try and develop as well as improve light weight security solutions to make the IoT Low Power devices more secure [6]. Below is figure 1 illustrating the IoT device life cycle from the manufacturing phase till the operational phase;

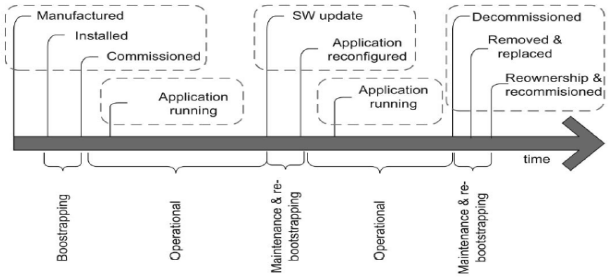


Figure 1: IoT device life cycle

Denial-of-Service attack, Evil Twin, Man-In-The-Middle as well as other attacks will be demonstrated in Section III.

III. Existing Hacking Techniques

Hackers use different techniques while performing the act of intruding. These techniques differ from hacker to hacker depending on the level of expertise the hacker possesses. This chapter will discuss existing hacking techniques as well as tools used to compromise the IoT communications networks with the focus on Kali Linux Platform. Moreover, it will dive into the hacker's mind to try and understand the compromise strategy followed.

3.1 Behaviors before attacking occurs

There are a series of phases an attacker follows when pre-planning his data violation act of hacking taking into consideration the lowest and the highest possibility of being detected as per figure 2 below;

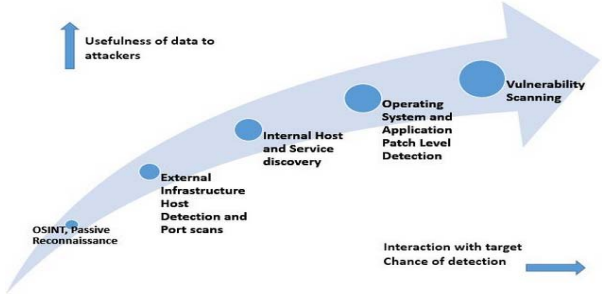


Figure 2: Interaction with target chance of detection

Reconnaissance is a first phase method used by the hackers to collect as much information as possible about the target they are aiming their intrusion at. There are two types of reconnaissance. The first type is called passive reconnaissance where the hacker collects information without directly engaging with the target using Open Source Intelligence (OSINT). This method uses a process of searching and collecting every available information related to the target. The OSINT method has two different approaches. The first approach is the offensive approach where the user collects information and prepare to attack a target, while the second approach is the defensive approach where the user collects information and prepare to defend a network. In this section, the focus will be on the offensive approach. Once the hacker collected the valuable information, a list can be created to be used during the hacking process. Some examples would be IP addresses or Domain Name Systems. The second type of reconnaissance is called active reconnaissance where the hacker engages directly with the target executing his/her active hacking techniques. This process poses more risk of being detected than passive reconnaissance. The hacker also uses Open Source Intelligence (OSINT).

As mentioned above, detection is a weakness in active reconnaissance. It is critical for attackers to be anonymous undetected during the hacking process. One of the methods used by the attackers is to change the mac address of their machines as a first step before starting any intrusion process. Once the mac address of the hacker's machine is changed, the attacker will use Stealth Scanning Strategies, to avoid being detected, where the hacker anonymizes himself/herself connecting to various unknown networks making it very difficult for the target to detect the intrusion act. One of the most well-known Stealth Scanning Strategies is connecting over the Tor Network which can create an IP spoofing scenario faking the source IP address after the hacker has modified the proxy chains used within the Tor network for anonymous communication purposes.

To resolve this risk and eliminate almost every possibility of being compromised, the hacker will configure his/her machine in a way where the attack platform (Kali) is sandboxed in a virtual machine and the Tor Network is being activated and Sandboxed via the Whonix Platform. The Whonix platform is a built-in platform which already has a virtual workstation and a Tor gateway which provides privacy, security and anonymity. This process will be tested in a lab demo.

The Offensive Intelligence approach, as mentioned, starts with collecting information for the aim of attacking a target using different tools. One of the most advanced tools used by hackers nowadays for reconnaissance purposes is called Maltego [9]. Maltego is an Information Gathering and a Social Engineering tool built in Kali Linux Platform and is used to collect information based on several categories that will not be all discussed in detail in this paper. NMAP is another very successful tool used by hackers. ZENMAP is the Graphical User Interface of NMAP which provides a more user-friendly interface. This tool is a Scanner and Network Exploration Tool.

A known method called Social Engineering is also used by many hackers for collecting valuable information about the target. This is a method used to deceive a human into releasing information the hacker can take advantage of. It is used in many ways whether by e-mail, phone call, direct conversation or any other form.

After discussing passive and active reconnaissance as well as social engineering, we will look at IoT devices in wireless environments. The focus in this section will be on IoT Wireless Networks. Performing passive and active reconnaissance on wireless networks might take some time and it requires several methods and tools. The hacker starts by preparing his/her machine and installing tools such as airmon, aerodump and aireplay. Once all the above has been installed, the hacker will be ready to use these tools for wireless sniffing, denial of service as well as intrusion detection which will be demonstrated in a lab environment [9].

The next topic to be discussed in this section is about a tool called Kismet. Kismet is a sniffer where the hacker can sniff the BSSIDs, IoT Devices "Stations", Channels, Tx, as well as the ESSIDs of wireless devices on a wireless network. Once the data is sniffed and analyzed, the hacker will choose an IoT Device as an example and will execute a Denial of Service attack eliminating the connectivity of this device to the network. Brute Force Attack is another form of attack where the hacker uses many combinations of letters and symbols to try and figure out the passwords / keys for wireless networks such as WPA. Every user that has a station such as a laptop computer must be authenticated before connecting to a wi-fi network via an access point. This authentication gets done between the user device and the access point through a process called four-way handshake where the client and the AP exchange packets confirming authentication of the client via the WPA key confirming the legitimacy of the user.

A Brute Force Attack can take advantage of such process by de-authenticating the connected machine to the AP and forcing it to authenticate again while sniffing the packets during the re-authenticating process allowing the hacker to crack these packets and get hold of the WPA key exchanged. Evil Twin is another attack method where the hacker sniffs the wireless environment first. Once ESSID of an IoT wireless device is detected and the hacker knows the name of the wireless device broadcasting, the attacker will launch a Denial of Service attack on the broadcasting IoT device putting it completely out of service and at the same time create a Rogue wireless device giving it the same name. This technique will trick the user who will lose internet connectivity and when trying to reconnect, he/she might reconnect by mistake to the rogue access point. This process will mis-lead the user to enter the IoT wifi key in a fake web interface created automatically when the user tried connecting to the rogue device. Once the user enters the key, the hacker will immediately get hold of the key and the rogue access point will terminate giving the user his original access back to the legitimate device without him/her noticing what just happened. One of the best tools in an Evil Twin attack is Fluxion which will be demonstrated live for

lab testing purposes. The hacker uses Fluxion to sniff the airwaves, launch a Denial of Service attack, capture a four-way handshake and create a rogue IoT device as the last step in the attack process. Another type of attacks is present and targets the Hyper Text Transfer Protocol (HTTP) and the Hyper Text Transfer Protocol Secure (HTTPs). This affects directly any IoT Device Web Interface where a Man-In-The-Middle attack takes place allowing the hacker to sniff information between the user and the IoT device. Websploit is an example of a hacking platform used by hackers for such attack type.

The Delivery is a second phase where the hacker has accumulated a sum of background useful info and moves towards selecting the hacking techniques and tools to be used to best compromise a system. The Compromise is a third phase where the attacker has already exploited the weakness in a system and achieved his set goal of gaining access. The Post-Exploit is a fourth phase where the attacker has already executed his plan and the damage has occurred. There exist several damaging scenarios where the outcome depends on the hacker's intentions and aim behind executing this act. [9]

IV. Intrusion Detection Systems

Every single node on any network must be protected to guarantee user information confidentiality, integrity and availability. For this reason, different devices were manufactured with the ability to protect networks perimeters and the nodes connected to it. These devices vary in functionality and are given different names such as Intrusion Detection Systems IDS, Intrusion Prevention Systems IPS or Firewalls.

IDSs are at the front edge of the network called perimeter. They are the first line of defense to combat malicious activity against the network. IDSs can be programmed manually where the network administrator loads specific set of rules to allow or deny certain activities. Based on these rules, the IDS will monitor the network behavior and makes sure every single activity which doesn't match the specified rule is detected. Once the activity is detected, an alarm will be triggered, and a notification will be released. IDSs do not have the capability to block attacks. Their main capabilities are limited to detection only [6].

If the network has only one IDS, this IDS device will be the centralized security platform with the weakness of being the central point of failure that might affect the entire network. If the network has several IDSs then it is better protected. If one of the IDSs failed, the network can still be monitored via the other IDS. However, this does not mean that the network is protected against Denial-of-Service attacks as an example. Think of the node on a network like a soldier in the battle field. When the soldier communication gets interrupted and he gets surrounded, the command center will lose track and the possibility of the soldier being killed or hijacked is extremely likely. The same applies to the node, IDSs fail in protecting against a denial-of-service attack. The hacker can disturb all forms of communication coming to the node, surround it and kill it or hijack it. Once the malicious node re-connects again, there is a high possibility the IDS won't even realize that is node is a rogue device.

A research done by group or researches and published under the title "Denial-of-Service detection in 6LoWPAN based Internet of Things" explains a network called the ebbit network where Suricata open source IDS was used with IDS-Probes to detect attacks such as Denial-of-Service attacks [6]. The Suricata platform is basically made of three components as mentioned in the study: the protocol decoder, the detection engine and the alert response unit.

Before performing the intrusion detection, the Suricata platform was pre-loaded with specific rules related to the attack to be performed for testing purposes where a certain threshold is considered the maximum before an alarm trigger. For this specific test, it was programmed on 30 packets per second. The testers performed an IPV6 UDP flooding attack from five different nodes against a node running 6LowPAN using an open source program called Contiki for IoT. The testers demonstrated how Suricata successfully detected the flooding attack when the packets transmitted exceeded the pre-loaded margin of 30 packets per second.

The mentioned research did not address the weaknesses of the Suricata system used and the dis-advantage such open source platform can have in a live commercial environment. A different research was done at the Naval Postgraduate School in California where weaknesses of Suricata appeared clearly during the test conducted. The study compared Suricata platform with Snort identifying the Suricata Platform as being the advanced modified new version of the snort platform. However, the study raised some major concerns when it comes to Suricata where some of the weaknesses in the platform is related to the false positives vs false negatives generated. Moreover, the study mentioned further weakness where the Suricata Software requires continues upgrades which makes it not suitable at all for production environments [7]. This makes Suricata with no added value to Wireless Industrial Sensor Networks WISN.

In section I, WSNs were discussed as a critical component in industrial environments. Cisco Meraki, the world leader in cyber security space, has developed one of the most advanced Intrusion Detection and Prevention system called Meraki. The Meraki technology can detect from any 802.11 devices the probe requests as well as the frames if the access point is within the Meraki network range. The same applies to mobile phones who also send probes trying to discover wireless active networks in the surrounding. Meraki has one of the most advanced secure management tunnels which is continuously communicating via a send/receive process that records statistics of devices available keeping the consumption bandwidth very low at 1 kbps. The Meraki perimeter access points can detect the signal strength of any available access point within a 100 feet range allowing the system to discover the actual physical location of the probing devices. This allows identification of hacker's locations in the case of rogue access points. The Meraki access point uses Wireshark discussed in the previous chapter to complete the probes sniffing process. The Meraki capability of detecting locations allows the collected data to be placed on different platforms in a form called "HeatMap" such as google earth to better show the location of specific users within a site as demonstrated in Figure 3 below:

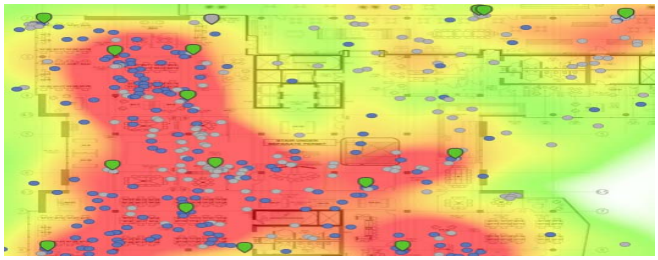


Figure 3: Meraki HeatMap

All data collected is channeled to a centralized management system which controls entire network. This system is in a form of a powerful device with a dashboard providing all details about the network. In this section, the Meraki centralized management system which will be discussed is known as the Meraki MX250 showed in Figure 4 below;

Figure 4: Meraki Centralized DashBoard

Meraki cyber security intrusion detection solution is one of the most advanced solutions recently deployed. It provides various advantages and added value services such as traffic shaping as shown in Figure 5 below;

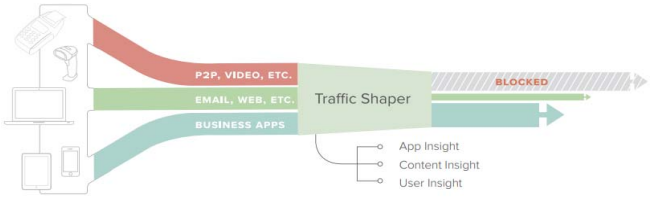


Figure 5: Meraki Traffic Shaper

The Meraki solution provides traffic shaping from the lowest level on the network at the IoT Low-Power Sensor all the way up to the Application Layer. The latest Meraki technology allows for deep traffic analysis leading to identifying the user behavior on a network. This was not enough, and deeper traffic analysis was needed leading to the development of a new classification scheme which provided the ability to create signatures on a dynamic basis based on host names and IP addresses. A more defensive approach from Meraki was to introduce Air Marshal solution where a Meraki Access Point responds to a Denial of Service or Evil Twin attack by sending back probing signals to jam the hacker device attempting the attack. A combination of Intrusion Detection as well as intrusion Prevention in one of the best most recent solutions [17].

V. Conclusion and Future Work

This paper summarized the recent development of the security issues relevant to IoT. From the overview, weaknesses expose even in the most advanced solutions. The main concern includes the effects of intrusion detection systems on overall network performance, because of the Air marshal process which enables the anti-intrusion systems support the network traffic and must deal with terminating the attack at the same time. Sending back probing packets to terminate a rogue access point or any other suspicious device at the same time performing network services has a downside effect on the overall network performance.

Our future work will recommending a better and more effective Intrusion Detection Scheme for IoT without the downside effect on network performance. A more effective Intrusion Detection Scheme will be developed by presenting an active in-depth-defense strategy and algorithm against these attacks relying on the intensive research that will be done in this field with the focus on 6LowPAN using penetration testing capabilities. The "Solution" will add value and be part of the most recent studies to combat cyber threats affecting the IoT infrastructure

References

[1] K. Carruthers, "Internet of Things and Beyond: Cyber-Physical Systems," *IEEE Internet of Things News Letter*. 10-May-2016.

[2] S. Li, L. D. Xu, and S. Zhao, "The internet of things: a survey," *Information Systems Frontiers*, vol. 17, no. 2, Apr., pp. 243-259, 2015.

[3] J. Gubbi, R. Buyya, S. Marusic and M Palaniswami, "Internet of things (IoT): A vision, architectural elements, and future directions," *Future Generation Computer Systems*, vol. 29, no. 7, Sep., pp. 1645-1660, 2013.

[4] O. Iova, P, Picco. T, Istomin, and C Kiraly, "RPL: The Routing Standard for the Internet of Things…Or Is It?," *IEEE Communications Magazine*, 16 Dec., pp. 16-22, 2016.

[5] R. Sahay, G. Geethakumari, and K. Modugu, "Attack Graph - based Vulnerability Assessment of Rank Property in RPL-6LOWPAN in IoT," In Proc. IEEE 4th World Forum on Internet of Things (WF-IoT) `02, 2018.

[6] P. Kasinathan, C. Pastrone, M. A. Spirito, and M. Vinkovits, "Denial-of-Service detection in 6LoWPAN based Internet of Things," In Proc. IEEE 9th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob) `10, 2013.

[7] E. Albin, and C. N. Rowe, "A Realistic Experimental Comparison of the Suricata and Snort Intrusion-Detection Systems," In Proc. IEEE 26th International Conference on Advanced Information Networking and Applications Workshops `03, 2012.

[8] S. Shin, T. Kwon, G. Jo, Y. Park, and H. Rhy, "An Experimental Study of Hierarchical Intrusion Detection for Wireless Industrial Sensor Networks," In Proc. IEEE Transactions on Industrial Informatics `09, 2010.

[9] V. K. Vijay, *Mastering Kali Linux for Advanced Penetration Testing*, 2nd ed. Reading, Packt Publishing Ltd, 2017. [E-book] Available: Packt e-book.

[10] A. Rao, N. Carreon, R. Lysecky, and J. Rozenblit, "Probabilistic Threat Detection for Risk Management in Cyber-Physical Medical Systems," *IEEE Software,* 25 Dec., pp. 38-43, 2017.

[11] C. Ionnou, and V. Vassiliou, "The Impact of Network Layer Attacks in Wireless Sensor Networks," In Proc. IEEE International Workshop on Secure Internet of Things (SIoT) `09, 2016.

[12] N. Japkowicz, and Y. Elovici, "Introduction to the Special Issue on Data Mining for Cybersecurity," *IEEE Intelligent Systems*, 11 Jun., pp. 3-4, 2018.

[13] C. Kolias, A. Stavrou, J. Voas, I. Bojanova, and R. Kuhn, "Learning Internet-of-Things Security Hands-On," *IEEE Security and Privacy*, 03 Feb., pp. 37-46, 2016.

[14] D. Ott E, "IoT Security through the Lens of Energy Efficiency: Energy as a First-order Security Consideration," In Proc. IEEE Cybersecurity Symposium (CYBERSEC) `04, 2016.

[15] Z. Ali Khan, and P. Herrman, "A Trust Based Distributed Intrusion Detection Mechanism for Internet of Things," In Proc. IEEE 31st International Conference on Advanced Information Networking and Applications (AINA) `05, 2017.

[16] W. Meng, "Intrusion Detection in the Era of IoT: Building Trust via Traffic Filtering and Sampling," *IEEE Computer Society*, 31 Jul., PP.36-43, 2018.

[17] Meraki, "Layer 7 Visibility and Control," Cisco Systems Inc., San Francisco, CA 94110, 2013. [Online]. Available: https://meraki.cisco.com/lib/pdf/meraki_whitepaper_layer_7_visibility_and_control.pdf

A Multi-Criteria Model for Users Trust Management in Ad-hoc Mobile Cloud Computing

Ayotuyi Tosin Akinola and Matthew Olusegun Adigun
Department of Computer Science
University of Zululand
KwaDlangezwa, South Africa
ruthertosin@gmail.com

***Abstract*—Ad-hoc Mobile Cloud (AMC) came up as a result of the need to alleviate the inadvertent and incessant connectivity challenges that are experienced in Mobile Cloud computing (MCC). Thus, mobile devices teamed up in groups to share their resources among one another which are majorly Web Services, Storage and Computing resources. However, potential participants in AMC often feel a level of insecurity as they envisaged loss of control over various personal and confidential data in their mobile devices which can occur in the course of sharing of resources. Therefore, AMC as a service provisioning paradigm for mobile device users has a need for an implementation of a trust management system to serve as a protection and guarantee of confidence for intended AMC participants. This study proposed a Multi-Criteria Trust Management system (MCTM) architecture for AMC to protect the interest of various participating mobile devices. This system provides an avenue to identify trustworthy mobile devices to whom they can carry out resource sharing and as well alert or notify them of any malicious act that could have happened. We carried out a simple evaluation procedure with a number of incorrect service matches in the course of responding to a query as well as comparing the level of precision attained by the proposed architecture with already existing research. The proposed system proves to give a maximum confidence to mobile users to show high interest in participating in the AMC system.**

Keywords—AMC, MCC, Multi-Criteria, Trust

I. INTRODUCTION

The cloud computing paradigm has been a distinguished platform for all forms of service provisioning to various forms of users in recent times. It has been a platform which ensures that all forms of resources can be turned into service being made readily available for any form of device to consume it [1]. Some typical and common examples of services that are available are Platform as a Service, Software as a Service and Infrastructure as a service. Others include Security, Antivirus and Database as a Service etc which are made available to various clients [2]. In general, any device that can make use of these various services on the cloud platform must be Internet-enabled to be able to connect to the cloud server. The Internet serves as the medium through which resources are shared with various devices such as personal computers and other forms of mobile devices.

Mobile Cloud Computing (MCC) is a special term used to describe all forms of mobile devices being connected to the cloud server for service consumption [3]–[5]. However, the challenge that is often experienced with various mobile devices with respect to network connectivity brings about the evolution of the platform that is independent of infrastructure called Ad-hoc Mobile Cloud (AMC) [6], [7]. AMC explains a scenario whereby various kinds of mobile devices open up their storage and computing resources for the benefit of sharing with other devices within the mobile cloud system. Various mobile users can enjoy this today as a result of vast improvement in the computing power of portable mobile devices in different forms which are available in the market. They come in various sizes and memory capacity which enables an efficient and effective performance with respect to offloading of data and sharing of various forms of services. They can now practically carry out the functions of Portable laptop and desktop computers conveniently with ease.

Considering the various advantages that come with AMC, among which include cost efficiency, quick deployment, fast and easy access to information, automatic service integration and with an almost unlimited storage enablement, it suffers from the possibility of mischievous mobile nodes as well as the lack of trust among the mobile cloud participants. Thus, there is a need for a mechanism that verifies the reliability, dependability as well as the vulnerability of each of the intending mobile devices. To this effect, this work proposed a model for users trust management system for AMC to ensure that users maintain a level of confidence when participating in an AMC system. The goal of this model is to ensure that no malicious mobile node is able to carry out any undesirable function or operation within the mobile cloud after joining the AMC system. A multi-criteria event-log system is deployed to enhance filtering of malicious nodes and severing them from infiltrating other mobile nodes to either confiscate their personal data or access confidential information without permission.

Several approaches have been used in the cloud computing paradigm and MCC to ensure that trust was inculcated into the infrastructure systems, however these approaches cannot be directly implemented into the AMC due to its infrastructure less nature that needs to be designed to take care of certain features such as the mobility and incessant disconnecting of mobile devices. Some of these works shall be discussed under Section II of the article. However, we simulate our proposed model for trust management within the AMC by using a mobile device simulator called Peer-to-Peer Simulator [6], [8]. The simulator provides a platform upon which transactions between various mobile devices that act as clients and providers can interact. Thus, further implementations on P2P simulator can be carried out to depict that exact performance that is required for the trust management system deployment.

This research work seeks to address the mobile users' lack of trust in the likelihood of occurrence of malicious mobile nodes which can result into a serious attack to a harmless participant within the AMC system. The remainder of this paper is organized as follows: Section II provides the

underlying literature review for this write-up and Section III presents the proposed model for multi-criteria trust management in AMC platform. In Section IV, we discussed the P2P trust simulator and its enhancing component for the model deployment. The model implementation was discussed in section V. Section VI provides the performance evaluation of the model in comparison to two other related works while section VII contains conclusion and proposed future work.

II. Literature Review

Trust management systems are daunting tasks to be deployed to any service provisioning platform due to various criteria that need to be considered against the possibility of entertaining any form of lope-hole or weakness [6]. Such platforms are expected to act as a check-up for all forms of illegal operations that might be initiated by any participating mobile nodes within the cloud system. Considering the AMC as a coexistence of both business and volunteer cloud users, the need for a trust management system becomes highly imperative [9]. Several research articles have described various approaches to trust management system and thus have been able to achieve a high level of protection for the various service consumers. Among these include the work of UI Haq and others in [10]. The authors proposed an SLA (service level agreement) validation schemes which compelled the service providers to provide a written compensation for violating the agreed term of the contract. Sometimes, compensation in terms of service credits is provided to the users for failure to deliver a certain level of service to the clients. However, it has been a challenge to harmonize a standardize SLA validation to be used by all service providers and also its full implementation has been a challenge in practice.

The work in [11] built more on the SLA validation technique in ensuring users' trust is improved. The work proposed a certificate and reputation-based model as part of the integration to the SLA validation framework. The work implemented a hybrid distributed trust model which makes use of Public Key Infrastructure (PKI) to prevent SLA abuses through recognizing susceptible malicious services during service provisioning. In a similar manner, trust enabled web service selection was modeled using both the domain characteristic of service-oriented computing as well as trust relation characteristics [12]. A new time related backward cloud generation algorithm was developed to generate the characteristics of cloud models in a numerical format thus serving as subjective options to various cloud clients. However, the two approaches described above do not include feedback from various cloud users and did not take into consideration the performance assessments from subscribing third party users.

In cloud computing paradigm, the use of trust and reputation system have been deployed to enhance better performance in web service selection. The research report in [13] depicted the number of criteria that were needed to be considered when selecting the optimal web service in the cloud. They were able to itemize various qualities which are more than the common QoS of Web Services which are needed to be considered for better performance. In addition, the work proposed a topology for classifying both reputation and trust systems via personalized or global, decentralized or centralized as well as resource or person. This work paved the way for understanding of the various parameters that will be of important value when designing the trust management model for this particular research report.

Considering the infrastructure-less platforms such as MANET and Ad-hoc mobile cloud systems, trust management model becomes a greater challenge as mobile nodes interact with one another though not residing at a particular spot [8], [14], [15]. Their transient nature coupled with the limited memory capacity with higher processing power usually affects the level of implementation and deployment that can be carried out. The work of Ries and Heinemann in [16] provided a new methodology for estimating the level of trustworthiness of entities along with an approach to weigh various recommendations as well as filtering appropriate trust from user responses which yielded a trust model that was called CertainTrust. The approach in this work is very impressive however the modalities for deciphering the trust from direct experience of the user becomes a limiting factor to the approach.

A private virtual infrastructure (PVI) was proposed as well in the work of [17] where a security architecture was used to enhance a trust model which shares the security mechanism that was proposed between the clients and service providers. Moreover, a research carried out by Thomas and Kalogeraki in [18] modeled a decentralized trust management middleware for peer-peer ad-hoc networks. The middleware enables the fate of a node to be dependent on the neighbouring node assessments which are often provided when needed. This is achieved through deploying a simple protocol which runs on top of the available infrastructure for message exchange. The middleware thus counts on the nonexistence of infrastructure to accomplish better trust in a fortified way.

In the same vein, the authors in [19] achieved a dynamic trust management mechanism using a cognitive cycle in distributed Cognitive Radio Networks. The secondary user's behaviour varies with the variation in various metric factors that was adaptively deployed to depict the present behavioural performance of each user. The evaluation of the work shows that the proposed mechanism enables more efficient and fairly management of the limited spectrum resources. However, this work was not within the context of Ad-hoc mobile clouds because we are not dealing with the primary and secondary users.

A more closely related work was the research carried out in [6] which uses an Eigen Trust algorithm to determine the reputation trust of each node and classified various nodes based on the derived reputation but there is no mechanism in place that can stop intending malicious nodes and reveal the intended actions that the malicious node wants to carry out.

A good trust management should not only alert the users when a malicious node is about to perform a mischievous action rather it depicts to the participating nodes how their confidential data or resources are being managed in the mobile cloud. It is with this view in mind that we proposed a multi-criteria trust management architecture for ad-hoc mobile cloud system to ensure users are aware of how their mobile resources are shared and how they are protected from malicious users. The awareness is derived from the alert system that notifies various users of the activities that go on around the mobile cloud. Once any illicit action or act is noticed, the alert mechanism is synchronized to notify the users.

III. Proposed Multi-Criteria Trust Management Platform

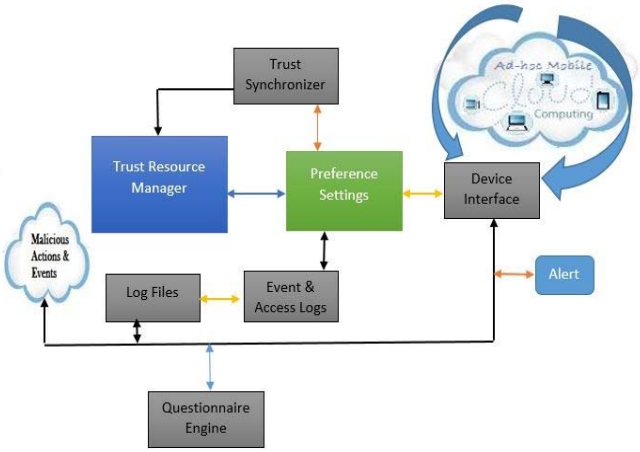


Figure 1: Multi-Criteria Trust Management Architecture Overview

The proposed multi-criteria trust management system for ad-hoc mobile cloud makes use of many components and criteria to ensure that users' interest are protected and that users can monitor every event that is carried out on their resources. In order to achieve the required performance that gives various users the opportunities to ascertain the use of their resources and the notification of intending malicious nodes intrusion, the following design must be critically considered and implemented as illustrated in the architecture in Fig .1. Thus, the elements of the multi-criteria design include the following:

A. *Dynamic Events definition*

This is taking care of by the preference settings module which enables the cloud users to specify events of interest to them. The preferences that each of the mobile users specified is kept on their record to be made available to them. They also have the control to deny the access to some services that they can render to other mobile nodes. Any infringement on the services that they are not ready to render will definitely raise a signal to alert the user of illicit or untrusted mobile nodes. However, the events of interest are made available to each of the participating nodes, provided that there are nodes whose preference settings allow the provisioning of such services.

B. *Denying unauthorized access*

Various level of access rights are provided for each individual nodes. The various permission will be different from one user to the other depending on the desire of the provider. E.g. read, write, create, delete and edit operations are some of the levels of access right that can be enabled and disabled with respect to a particular user. These permissions keep the cloud participants from any form of mobile and insider threats through keeping the track of what other mobile devices are doing.

C. *Prompt data encryption*

This is to be deployed to enhance that strictly confidential data are protected on the transmission to the intended destinations and that it does not stay long on the server so as to prevent unauthorized access. Various high-level encryption algorithms such as RSA, AES or Triple DES among others are

Figure 2: Catalogues of event logs on a mobile node.

typically deployable algorithms. However, AES was used in this particular work for this purpose.

D. *Event Automation*

The detection of a malicious node is one of the goals of the trust management system and that is carried out in an automated manner such that prompt response is sent to the concerned mobile user. Reporting of illicit events must be automated based on the filtering off of unpermitted actions with the trust synchroniser module assisting in keeping track of such events in case of future reoccurrence or a path to such a similar request which must be alerted abruptly.

The architecture overview was designed to speak to all the aforementioned points with respect to the trust management system. The Log files and the events with access logs are constantly checked to ensure that permitted activities were going on in the mobile cloud that does not predisposes participating mobile nodes to any form of danger. Fig. 2 shows a typical event log files shown on participating mobile devices in the ad-hoc mobile cloud system. These log files assist the trust management system in the following ways (1) enhance the users to have a clear picture of what transpired with their resources, (2) detect where the files are kept, (3) enable the clients to understand the manner of protection that is rendered thereby giving them some levels of confidence and (4) the beneficiary of the files that are available on the AMC. Moreover, the log files provide information through the time zone and stamp lines which help to detect where the intruding mobile device was when the attempt was being made.

IV. P2P Simulator and its Components

This P2P simulator [20] is good at testing the performance of a proposed systems in a peer to peer connections. Its mode of operation usually occurs in two stages according to [6] where it generates a trace file that contains the number of nodes with their reputation values. The values of the generated reputations were based on the computations from the parameters derived from the Questionnaire Engine section that was implemented in the proposed architecture. The second step makes use of the trace file that is generated to derive statistical outputs after applying a trust algorithm as shown in Fig .3.

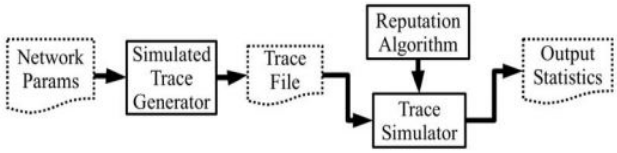


Figure 3: Overview of the evaluation architecture [20]

The evaluation architecture in Fig .3 uses EigenTrust algorithm to determine the reputation value for each of the participating nodes. The values that were determined in the first simulation influences subsequent values that are determined in the next round. Under each round of events, a new_trust_threshold value is determined and this is synchronised with the Trust Manager to influence future events.

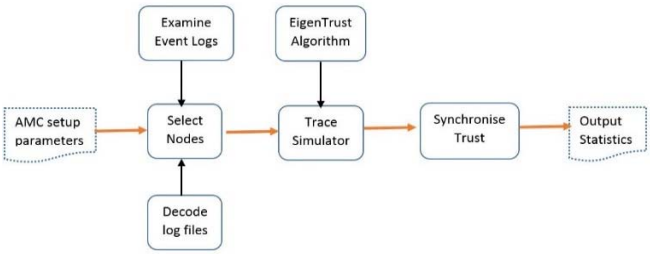


Figure 4: Overview of the evaluation architecture with multi-criteria enablements

The new enablement allows the use of event logs and log files to depict the proper assessment of the node behaviours/reputation. The trust response manager contains private information accessory and files log analysers which usually take care of the log files and event logs. Its purpose is to ensure that all the preferences and privacy that pertains to the individual mobile nodes are securely taken into consideration. Considering the feedback responses through the questionnaire engine alongside with the log files and events with access logs serve as the major source of information that enables the functionality of the Trust Resource Manager. We classified the various nodes into the following 5 groups as depicted in Table 1.

The various categories were assigned a measure of the level of trust based on the user experience as 0.2 from the least to the peak with a value of 1. A similarity computation and *prediction creation* steps were carried out as explained in our previous article [21]. The result from the prediction creation is used to determine the most reliable and best performing mobile node to attend to the query that was issued in the AMC. Thus, the trust resource manager takes up the responsibility of analysing all the log files and critical examination of the event logs to report any form of error, abrupt status changes, software bugging and other forms of hacking through triggering the alert mode. This alert mode allows the targeted mobile nodes to decline or deny such malicious event from being executed.

V. Model Implementation

The model followed the described flow process in Fig. 3 in the same manner as used in the article [20]. The statistical output from Fig. 3 brings about the triggering effect that is carried out by the trust resource manager. The preference settings as well as the Questionnaire Engine works together to bring about the selection of better nodes during service query

TABLE I
Node Categorizations

SN	User Types	Representation	Symbols
1	Pre-trusted nodes	Most Credible	R_5=1
2	Trusted Nodes	More Credible	R_4=0.8
3	Good Nodes	Credible	R_3=0.6
4	Malicious Nodes	Less Credible	R_2=0.4
5	Undetected Nodes	Least Credible	R_1=0.2

within the AMC. Thus, both criteria work synchronously to enhance a stable and effective trust management mechanism for AMC systems. The P2P simulator enhances the derivation of new_trust_threshold value with each round of simulation process according to equation 1.

$$t = (l \times 0.2) + (s \times 0.2) + (r \times 0.15) + (c \times 0.15) + (a \times 0.3) \quad (1)$$

Where l, s, r, a and c represented latency, security, response time, cost and availability respectively of a particular client within the AMC. The derived value of t gives the new_trust_threshold with respect to a particular AMC setup. Looking at various parameters that are considered here, the relative importance of the parameter is shown by the magnitude of the allotted multiplicative product assigned to it to influence the importance of the parameter. A comparative or similar value of "t" and the output of *prediction creation* step being highest as the value gets closer to 1 affirms a trusted mobile node that can be allowed to carry out resource sharing within the system. This process clearly made it difficult for a malicious node to scale through in MCTM architecture.

VI. Performance Evaluation

This section tests the performance of the MCTM system for efficiency and user monitoring device usage. The goal here is to evaluate how the proposed MCTM performs incomparable to other research outputs. The experiments were conducted using over 5000 mobile nodes. An Intel core i7-4500U processor with 2.40 GHz CPU speed and 8GB RAM Asus machine was used for the experiments. Over 5000 different nodes were uniformly dispersed within the setup, and about 200 random nodes were used to test how the system will perform under fewer nodes, thus calculating the mean results of the metrics from 10 different measurements/runs.

We denoted some nodes in different categories as stated under Table 1, which were introduced into the network. Before the feedback system will begin to function appropriately, we concur to the fact that the malicious nodes can only be deceptive once after which its proper categorization will be determined and enlisted with the MCTM system. Our evaluation process compares the performance of PlanetCloud [22] and TMC [6] research work with MCTM system. Both [22] and [6] relied on the use of spatio-temporal calendar to identify good behaving, pre-trusted and bad behaving mobile agents. The 200 nodes were categorised into a group of 40 for each categorization user type to determine the mean number of incorrect matches that will be generated. Furthermore, we determined the performance of the system as well with respect

A. *The quantity of correct matches with varying percentage of trusted nodes and overall nodes.*

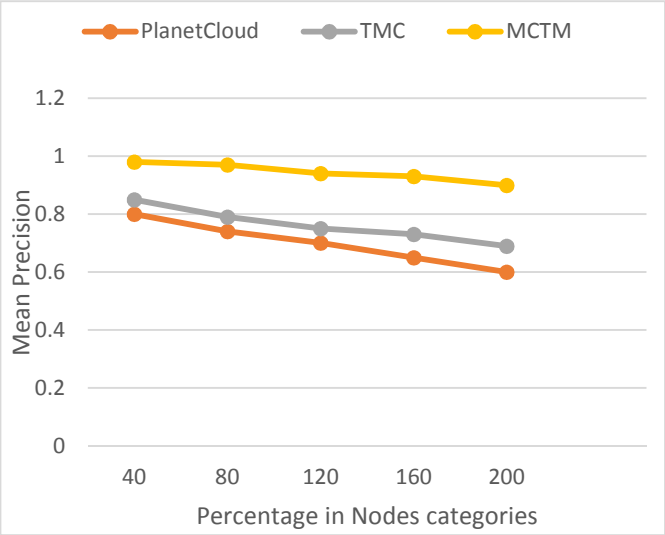


Figure 5: Quantity of true matches for different percentage of trusted nodes.

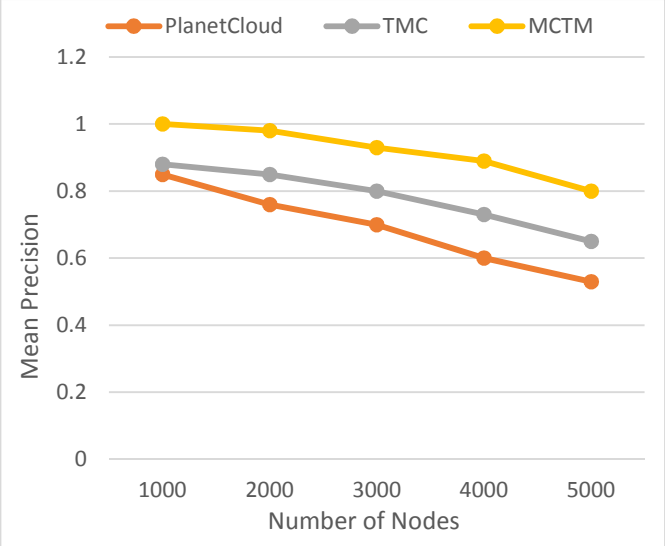


Figure 6: Quantity of true matches for different number of overall nodes.

to the mean precision when quantity of correct matches was considered with a percentage of categorized 200 nodes alongside with the the overall number of nodes. The overall number of nodes experiments has each category with 1000 nodes each assigned to them.

Fig. 5 shows the level of precision produced by different approaches with respect to the number of mobile nodes that were able to provide what was requested within the system. The recorded mean precision determines the level of accuracy of the proposed system under the case study for the experimental setup. The MCTM provides a higher level of precision compared with another system in this particular experiment, with the highest precision being approximated to 1. Only a few nodes were unable to provide the requested query while using the MCTM platform thus, making the level of its precision higher within the provided trusted nodes. However, a little deviation in MCTM shows nodes with almost similar

B. *Incorrect matches with varying percentage of trusted nodes and overall nodes.*

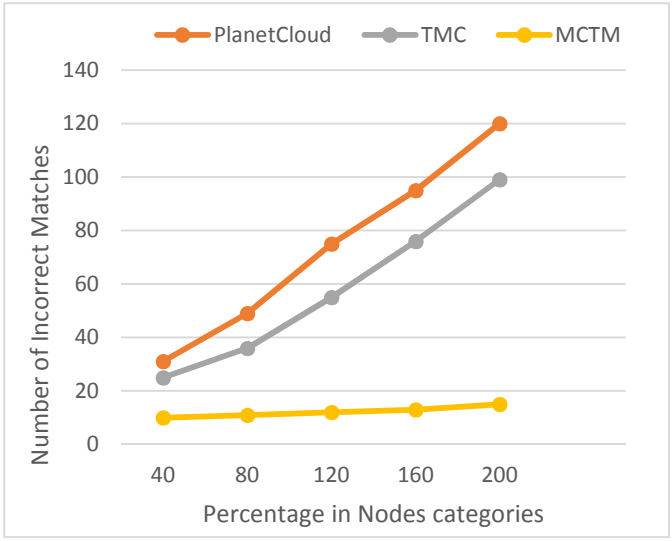


Figure 7: Incorrect matches with respect to varying percentage of trusted nodes.

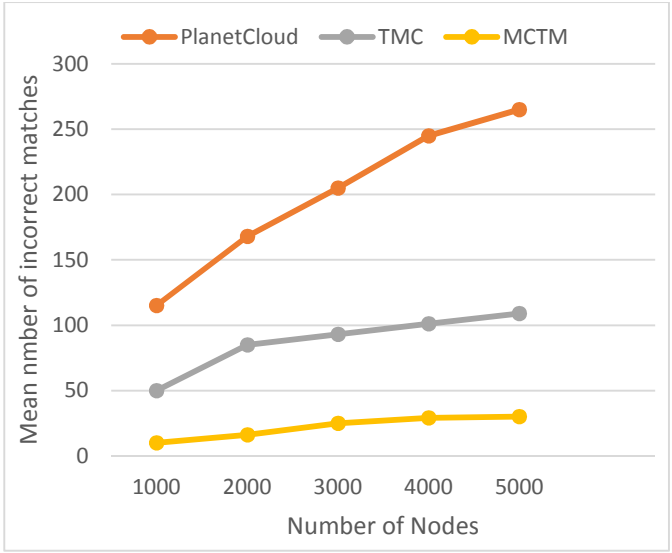


Figure 8: Incorrect matches with respect to varying number of overall nodes.

resources. However, it is counted as unmatched by the system.

Fig. 6 as well depicted the performance of the different approaches when the overall number of nodes were utilized. In this case, a total of 5000 nodes sample were considered and the mean precision is as shown in Figure 6. The performance depicted also shows a similar result as the precision level was steep with a larger volume of mobile nodes within the cloud system. The Multi-Criteria Trust Management proposed performs better with several node categorizations as shown in Table 1 by ensuring that only trusted nodes can operate within the cloud system. The precision performance of the MCTM in both experiments above was due to multi-criteria implantations which include event log, log files, feedback rating system as well as selective preference settings thus, increasing the level of precision attained by MCTM more than other approaches.

Under the experiment shown in Fig. 7, we try to determine the number of incorrect matches that were experienced in the course of evaluating the proposed management system in our

work compared with other approaches. With the various categories of mobile nodes in the network, we first determined the number of unmatched request in a 200 sample of trusted nodes. This test hangs on the ability of the system to ensure that trusted nodes were the only ones participating in resource sharing. Giving the fact that only 200 nodes were the number of those who can be trusted, then we determined the effect of this on the number of incorrect matches that find their way through to cause an incorrect matching in the system.

The results show a lower number of incorrect matches in MCTM for this particular experiment. While the number of incorrect matches increase with other approaches, it decreases in our approach due to the fact that our system platform critically examines the log files and event logs of each mobile devices to ascertain their peculiarities before the other mobile node can respond to such query or request. The effect of larger size sample was clearly displayed with the result of the experiment in Fig. 8. The number of incorrect matches increases as well, however, the MCTM shows a better comparable performance with respect to other solutions provided in ensuring that malicious nodes were denied access rights to express themselves as trusted nodes.

VII. Conclusion and Future Work

We proposed the Multi-Criteria Trust Management (MCTM) architecture to enhance the eradication of untrusted and malicious mobile nodes from accessing or corrupting the files of genuine users. The proposed system uses more than one criteria for inhibiting the malicious users unlike other approaches which use one or two. Our proposed system attained a better precision level and very low incorrect matches during the evaluation process thus, proving a better platform for inhibiting the effect of malicious nodes in the AMC. We also made the users to be aware of various processes that go on when they are connected to the mobile cloud. The users also were alerted whenever any malicious node make any undesirable move. Our future work will be to ensure that mobile node without the requested query are not allowed to attend to the request in the absence of none that has the requested query. We also envisioned to implement a prototype experimentation of this proposed architecture so as to evaluate how users experience look like when our work is deployed. We envisaged a prototype implementation via a test bed will provide us with real-world assessment of our proposed architecture.

Acknowledgment

The authors acknowledge the funds received from the industry partners: Telkom SA Ltd, South Africa in support of this research.

References

[1] A. Akinola, M. Adigun, and A. Akingbesote, "Qos-Aware Single Service Selection Mechanism for Ad-Hoc Mobile Cloud Computing," in *Proceedings on International Conference on Computing, Communication and Security (ICCCS) IEEE, Mauritius*, 2015, pp. 281–286.

[2] L. Wu, S. K. Garg, and R. Buyya, "SLA-Based Resource Allocation for Software as a Service Provider (SaaS) in Cloud Computing Environments," in *11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, 2011, pp. 195–204.

[3] H. T. Dinh, C. Lee, D. Niyato, and P. Wang, "A survey of mobile cloud computing : architecture , applications , and approaches," *Wirel. Commun. Mob. Comput.*, vol. 13, no. 18, pp. 1587–1611, 2013.

[4] A. Aissioui, A. Ksentini, A. M. Gueroui, and T. Taleb, "Toward Elastic Distributed SDN/NFV Controller for 5G Mobile Cloud Management Systems," *IEEE Access Recent Adv. Softw. Defin. Netw.*, vol. 3, no. 10, pp. 2055–2064, 2015.

[5] J. Shuja, K. Bilal, S. A. Mandani, O. Mazliza, R. Ranjan, P. Balaji, and S. U. Khan, "Survey of Techniques and Architectures for Designing Energy Efficient Data Centers," *IEEE Syst. J.*, vol. 10, no. 2, pp. 507–519, 2016.

[6] A. Hammam and S. Senbel, "A trust management system for ad-hoc mobile clouds," in *Intelligent Systems Reference Library*, 2014, vol. 70, pp. 519–539.

[7] A. T. Akinola, M. O. Adigun, and P. Mudali, "A Federated Web Service Selection Approach for Ad-Hoc Mobile Cloud Computing," in *19th Southern African Telecommunication Networks and Applications Conference (SATNAC) Frankhurt, South Africa*, 2016, pp. 386–373.

[8] C. Biao and L. Zhen, "Simpler Trust Simulator Model for Structured P2P Network," in *2nd International Conference on Industrial and Information Systems*, 2010, pp. 83–86.

[9] S. M. Habib, S. Ries, M. Muhlhauser, and T. U. Darmstadt, "Towards a Trust Management System for Cloud Computing," in *10th International Conference on Trust, Security and Privacy in Computing and Communications IEEE*, 2011, pp. 933–939.

[10] I. UI Haq, I. Brandic, and E. Schikuta, "Sla validation in layered cloud infrastructures," in *7th International Workshop, Economics of Grids, Clouds, Systems and Services, (GECON 2010), Ischia, Italy*, 2010, pp. 153–164.

[11] I. UI Haq, R. Alnemr, A. Paschke, and C. Meinel, "Distributed trust management for validating sla choreographies," in *Grids and Service-Oriented Architectures for Service Level Agreements Springer US,* 2010, pp. 45–55.

[12] S. Wang, L. Zhang, S. Wang, and X. Qiu, "A Cloud-Based Trust Model for Evaluating Quality of Web Services," *J. Comput. Sci. Technol.*, vol. 25, no. 6, pp. 1130–1142, 2010.

[13] Y. Wang and J. Vassileva, "A Review on Trust and Reputation for Web Service Selection," in *27th International Conference on Distributed Computing Systems Workshops*, 2007, pp. 3–8.

[14] F. Wenjuan, Y. Shanlin, P. Harry, and P. Jun, "A Multi-dimensional trust-aware cloud service selection mechanism based on Evidential Reasoning Approach," *Int. J. Autom. Comput.*, vol. 3, no. 5, pp. 1–12, 2013.

[15] S. Pearson, "Privacy, security and trust in cloud computing," *Priv. Secur. Cloud Comput.*, vol. 1, no. 1, pp. 3–42, 2013.

[16] S. Ries and A. Heinemann, "Analyzing the robustness of CertainTrust," in *2nd Joint iTrust and PST Conference on Privacy, Trust Management and Security, Springer*, 2008, vol. 10, pp. 51–67.

[17] F. J. Krautheim, "Private virtual infrastructure for cloud computing," in *HotCloud'09 Proceedings of the 2009 conference on Hot topics in cloud computing, San Diego, California: USENIX Association Berkeley, CA, USA*, 2009, pp. 1–5.

[18] T. Repantis and V. Kalogeraki, "Decentralized trust management for ad-hoc peer-to-peer networks," in *Proceedings of the 4th international workshop on Middleware for Pervasive and Ad-Hoc Computing (MPAC 2006) ACM*, 2006, pp. 1–6.

[19] Z. Long, L. Zhan, and G. Chen, "A Dynamic Trust Management Mechanism in Distributed CRNs," in *12th International Conference on Computational Intelligence and Security*, 2016, pp. 237–242.

[20] A. West, S. Kannan, I. Lee, and O. Sokolsky, "An evaluation framework for reputation management systems," in *Scholarly Commons as a book,* 2009, pp. 282–308.

[21] A. T. Akinola and M. O. Adigun, "Feedback-Based Service Selection in Ad-Hoc Mobile Cloud Computing," in *3rd International Conference on Advances in Computing and Communication Engineering (IEEE-ICACCE-16). Durban*, 2016, pp. 72–77.

[22] A. Khalifa and M. Eltoweissy, "A global resource positioning system for ubiquitous clouds," in *2012 International Conference on Innovations in Information Technology, (IIT 2012)*, 2012, pp. 145–150.

A Hybrid Recommendation Technique for Big Data Systems

Chitra Nundlall
Department of ICT
University of Mauritius
Réduit, Mauritius
shwetanundlall@gmail.com

Gopal Sohun
Department of ICT
University of Mauritius
Réduit, Mauritius
gopalsohun@gmail.com

Soulakshmee Devi Nagowah
Dept. of Software and Information Systems
University of Mauritius
Réduit, Mauritius
s.ghurbhurrun@uom.ac.mu

***Abstract*—Recommender systems are engines that recommend new items to users by analyzing their preferences. The web contains a large amount of information in the form of ratings, reviews, feedback on items and other unstructured data. These details are extracted to get meaningful information of users. Collaborative filtering and content-based filtering are two common approaches being used to make recommendations. The paper aims to introduce a hybrid recommendation technique for Big Data Systems. The approach combines collaborative and content-based filtering techniques to recommend items that a user would most likely prefer. It additionally uses items ranking and classification technique for recommending the items. Moreover, social media opinion mining is added as a top-up to derive user sentiments from user's posts and become knowledgeable about users' tastes hidden within social media. A prototype has been implemented and evaluated based on the recommendation techniques.**

Keywords—hybrid item recommender, social media, collaborative filtering, content-based filtering, sentiment analysis

I. INTRODUCTION

The evolution of latest technologies and the growth of cloud computing and Internet of Things (IoT) have promoted the sharp growth of data [1]. The data is proliferating in such a way that it is becoming time consuming to look for desirable items on the web [2]. Moreover, having smartphones or other digital devices in their possession, people are getting more connected to social media sites and a lot of blogs inundating them with user sentiments. They are less reluctant to post their feedback, reviews and opinions about any item (grocery, books, products) using their smartphones. This gives rise to a massive amount of unstructured data and semi-structured data combined with structured data called Big Data on the web. Big Data is characterized into 5 V's which are Volume (large amount of data), Velocity (speed at which data are being generated), Variety (different types of data), Veracity (accuracy of data) and Value [3], [4], [5]. Big Data can be used as the basis for determining user preferences [1],[4]. By extracting and analysing Big Data, organizations can gain insights about their customers and eventually provide them with personalized recommendations [6]. It is often the case that surfers visit websites but they hardly buy items. This is mostly because they are unable to make a choice out of the numerous items available in front of them. Being in a busy world, recommender systems have become indispensable for making decisions for users within seconds. Despite their success in the industry, recommender systems entail several problems [7], [8]. For some, the sparseness of the user-item matrix affects the recommendation quality while others ignore the correlation among users' tastes [7]. The data being used sometimes does not match with the user's interest correctly which then provide less accurate and personalized recommendations. Other issues faced by existing recommender systems are: insufficient amount of data, changing data and changing user preferences [9]. The volume of data available to recommender systems is forcing a total re-evaluation of the methods used to compute predictions [12]. These data cannot be used directly in recommender systems. They should be filtered and transformed as per the requirements [10]. The paper therefore aims to propose a hybrid recommendation technique for Big Data systems. The rest of the paper is structured as follows: Related work is discussed in section II. The proposed recommender system and the architectural design is described in section III. Section IV describes the implementation of the system. Finally, conclusion and future works are discussed in section V.

II. RELATED WORK

Content-Based Filtering (CBF) and Collaborative Filtering (CF) are techniques commonly adopted recommendation by recommender systems. Content-based recommendations are based on customer's profile that keep track of previously viewed items [11]. A system using CBF then suggests items almost similar to the previous ones viewed. For the system to recommend items, the user needs to rate a number of items so that the user's interest can be understood. It builds a matrix of items to other items and calculates similarity based on user rating. Items with the highest similarity score are the most similar [14],[18],[20]. CBF eliminates the no first rater problem. It uses explicit keywords from user profiles and attributes of the items that they share to make predictions [14]. The similarity between the item attributes and the user profiles are calculated in textual domains, thus eliminating a sparse relationship between user profiles and items [25]. However, in case content description is low or unavailable, prediction becomes less accurate. The algorithm often lags in identifying user tastes and opinions, resulting in a restriction for highly subjective recommendations [18],[21]. On the other hand, CF uses the preferences of multiple users which have related ratings, as feedback information to make recommendations [11]. An average rating is made out of all the items rating made by different users and then the item is proposed. New items are ignored until an acceptable amount of ratings are obtained. It uses a matrix which depends on the past interactions of the users, for example, the rows of the matrix are composed of the user and the columns are composed of

the items. The corresponding entries in the matrix are simply the rating that each user give for each item. It applies weights based on other user's preference. Items with the highest recommendation score are the most relevant to a user [18], [20]. The positive side behind CF is that the algorithm is not dependent on the content of the items when generating recommendations [14]. It only considers user profiles and provides the flexibility to filter out items based on its quality and/or a user's taste. Furthermore, it can provide unexpected good recommendations [25]. Nevertheless, it has some downsides, when a new user or new item is introduced, they usually have no past ratings information associated with them, which makes it difficult for the algorithm to make recommendations. The major bottleneck is starting afresh, technically referred to as the cold start problem. This limitation prevents or leads to less efficient (users with fewer rating cannot be placed in a good neighbourhood) predictions [13]. Enough ratings are required to make a sufficiently good recommendation [25]. Both CF and CBF techniques have their strengths and shortcomings. The disadvantage of one technique is the advantage of the other technique and vice-versa.

Other techniques like sentiment analysis can be considered when making recommendations to users. Sentiment analysis is the process of extracting meaningful information from social websites such as Facebook and twitter where users express their emotions and opinions in the form of text and emoticons as comments and postings [27]. Facebook and Twitter data are unstructured and usually contains many dirty words so it needs to be filtered and for that it requires to undergo a deeper pre-processing and manipulation to extract high quality information. Using suitable tools such machine learning [28], topic based text stream analysis technique [17], natural language processing techniques [19], opinion mining and sentiment analysis can be performed to analyse data like, for example, customers' product reviews.

Some recommendation systems are described as follows.

A. System 1: Hotel Recommendation System Using Hadoop Framework [24]

The Hotel Recommendation System makes use of a review based service recommendation method to make personalized recommendations. The CF algorithm is used to perform recommendations for a user. The system operates in such a way that users give their reviews. Stopwords are removed to get keywords. The keywords are compared with a dataset and the matching keywords are obtained. A ranking dictionary is also formed with ranking values from -1 to +1. Sentiment Analysis is then performed. Hadoop framework is used to run the system and Map-Reduce together with JAVA is used to reduce number of similar keywords.

B. System 2: Personalized Recommendation System [28]

Personalized Recommendation System is a recommendation engine that can recommend books to users based on their interest and features of the books. It makes use of both CF and CBF approaches to perform recommendations. The system takes the ratings of the user as input together with the user profile. The dataset is a set of books. The tool that is used to analyse the big data is Hadoop. JAVA and MapReduce are used to create the application. When a new user wants to log in, he must create an account in the system where a list of books is given and the user is asked to rate the books so that the system can analyse the interest of the user and predict other books to the user. Other tools such as putty and WinSCP are used to run Hadoop in the Hortonworks sandbox environment that consists of Hadoop components. Region aggregation, an interesting feature is implemented that makes an average of the ratings from different regions and perform predictions.

C. System 3: Big Data Analysis: Recommendation System with Hadoop Framework [10]

The system is a recommendation system that takes into consideration a person's likes and ratings left for multiple items, and identifies predictions for that person. It uses various forms of data originating from websites such as ratings, reviews and comments. The system uses the Hadoop framework for managing, storing and retrieving the large amounts of data obtained. The Hadoop framework interfaces with Mahout to analyse and generate insights from the data.

D. System 4: Docear's Research Paper Recommender System [16]

Docear is an open-source desktop academic literature suite written in JAVA, used to organize references and PDFs. It has a recommender system also written in JAVA and runs on web servers. It uses a RESTful Web Service to communicate between the desktop software and the servers. The recommendation engine creates new user models and recommendations whenever new mind-maps are uploaded to the server or after recommendations have been delivered to the user. Stereotype recommendations and content-based filtering (CBF) have been used. Apache Lucene is used to match user models and recommendation candidates. Once the recommendation is created, it is stored in the recommendation database (MySQL and Hibernate).

E. System 5: ShopZoo: A Grocery Mobile Application with Smart Recommendation System [15]

ShopZoo is a cross platform application that allows users to place orders of grocery, stationery and electronics items. The application collects a mixture of detailed customer reviews and ratings of items. It applies an item-to-item collaborative technique to display a list of shops based on a customer's geographic location and the store's product availability. The item-to-item collaborative technique works by finding users with similar tastes using the ratings supplied to make recommendations for a similar user. The technique uses a cosine similarity measure used to calculate similarity between items.

The systems are further categorised in terms of recommendation techniques (RT), data sources (DS) and environment (ENV.) in Table 1. While analyzing the five existing recommendation systems, it can be observed that not all systems take into consideration aspects of Big Data when giving their recommendations. Only [24], [10] and [28] use Big Data. These three systems do not, however, capture data from all the three data sources namely *User Profile*, *Open Source* and *Dataset*. [10] and [28] use CF and CBF. [24] uses CF and sentiment analysis. None of the five systems make use of all the three methodologies namely CF, CBF and sentiment analysis. Additionally, most of the systems are desktop applications. Nowadays, people prefer to adopt mobile applications which can be used anywhere and anytime.

TABLE 1. COMPARISON OF EXISTING SYSTEMS

	SYSTEM	1	2	3	4	5
RF	CF	✓	✓	✓	×	✓
	CBF	×	✓	✓	✓	×
	Sentiment analysis	✓	×	×	×	×
DS	User Profile	×	✓	×	×	✓
	Open Source	×	×	✓	✓	×
	Dataset	✓	✓	✓	✓	×
	Big Data	✓	✓	✓	×	×
ENV.	Desktop	✓	✓	✓	✓	×
	Mobile	×	×	×	×	✓

III. PROPOSED SYSTEM AND ARCHITECTURAL DESIGN

Based on the different observations, the paper therefore proposes a hybrid system that combines three different techniques namely CF, CBF and sentiment analysis and captures data from the three data sources namely *User Profile*, *Open Source* and *Dataset.* A system overview is given in Fig. 1. Fig. 2 presents the architectural diagram that uses a three-layer software pattern to describe the various components to be used within the proposed system.

A. Data Source

The proposed system is a mobile friendly web application that makes use of data coming from three primary sources: User Profile, user ratings and social media (unstructured data). A new user registers on the application and is authenticated using a username and password. Each user creates a profile that contains the preferences explicitly stated by the user. The user provides ratings for particular items. Users additionally input their feedback and reviews in the form of tweets in Twitter. The system works in such a way that as soon as the tweets are added by users, they are being considered by the system to make proper recommendation. Furthermore, the system is experimented using a large number of books from a freely available Book-Crossing dataset [29] that contains structured data. The dataset consists of 278, 858 users of who have provided 1,149,780 ratings for 271,379 books. It comprises of three tables: BX-Users, BX-Books and BX-Book-Rating. The structured data is converted into a nested dictionary as the algorithms take a dictionary as input.

B. Pre-processing of Dataset

To adapt the dataset to the proposed system, pre-processing of the data is required. The data firstly need to be cleansed to remove redundancy information. Some data are merged together in order to resolve conflict. Data integration is performed. The raw data is then be transformed into an understandable format for the algorithms. Keywords are extracted to get meaningful information from unstructured data in the form of comments/tweets.

C. Search

The end user is able to search a vast majority of daily updated database of books, through the Google Books API. The very moment a query is made, a REST service interfaces with the Google API to dynamically retrieve book details in JSON format. Thereafter, upon the user's selection of a book, relevant data such as book details and category are extracted and stored on the local database.

D. Recommendation Engine

The core component is the Recommendation Engine and it is responsible for generating recommendations for users in the system. It takes explicit data fed to the system, such as the user preferences and ratings left for items, and then outputs a list of items that users might be interested in. The engine makes use of a hybrid approach, which combines the two subcomponents: CF module and CBF filtering module.

The CF module makes use of two sets of data namely a set of users and a set of items. The module identifies similarity between users to infer recommendation for each user. Ratings define the relationship between users and items. First, the module accepts the current user (U_a) ratings as input, that is, it uses ratings to infer recommendation for each user. Secondly, it identifies users similar to U_a. The similarity between U_a and U_b is determined using the Pearson's correlation coefficient [26]. The Pearson's correlation is the chosen correlation metric as it addresses issues where one user has provided a rating while another has not, by simply considering items where both users has provided a rating. Thirdly, the set of items is then ranked using K-Nearest Neighbour (KNN). Finally, the list of identified recommendations for U_b is ready.

The CBF module makes use of two sets of data: a set of users (U_a, U_b, U_c, ...) and a set of items with keywords describing the items (an example for movies: action, horror and others would be used as keywords). Firstly, the module accepts the current user (U_a) preferences (personal interest) as input, for example, some user (U_b) might be interested in a certain category of book such as science-fiction. Secondly, it identifies items that are similar to items previously rated by the current user. For example, if U_a rates three items falling in the category of science-fiction, the module infers that U_a is interested in that category. Thirdly, KNN initially determines the proximity between each point from the dataset. Based on the k-neighbour set, it identifies the top closest points. Ultimately, it makes use of a voting strategy to count and insert most points to the class label of test dataset.

E. Sentiment Analysis

The system extracts sentiments from tweets using Twitter Search API. The API is used for requesting the tweets for a specific user. These tweets normally contain data about user ratings and preferences. The data is then stored in the database. From the database, the tweets are fed to the natural language processor whereby the tweet information is extracted and filtered based on a dictionary. Irrelevant tweets that contain words not useful are discarded. The results is then classified as either positive or negative depending on which score is dominant. The

sentiment score of each tweet is stored in a column in the table. The system then determines the genre of the book that the user is mentioning in his tweets as well as the book name. A twitter page (@worldbooksPage) has been explicitly set up, to allow users to post reviews concerning books of their interest, leading to creation of a large amount of unstructured data. An example of a tweet which is fed to the natural language processor and filtered based on a dictionary is "*OMG!! That horror book… damn interesting….xoxo*". By extracting the words 'horror', 'interesting', the results are classified as positive, that is, the user likes horror books. The book category is obtained and preference as well.

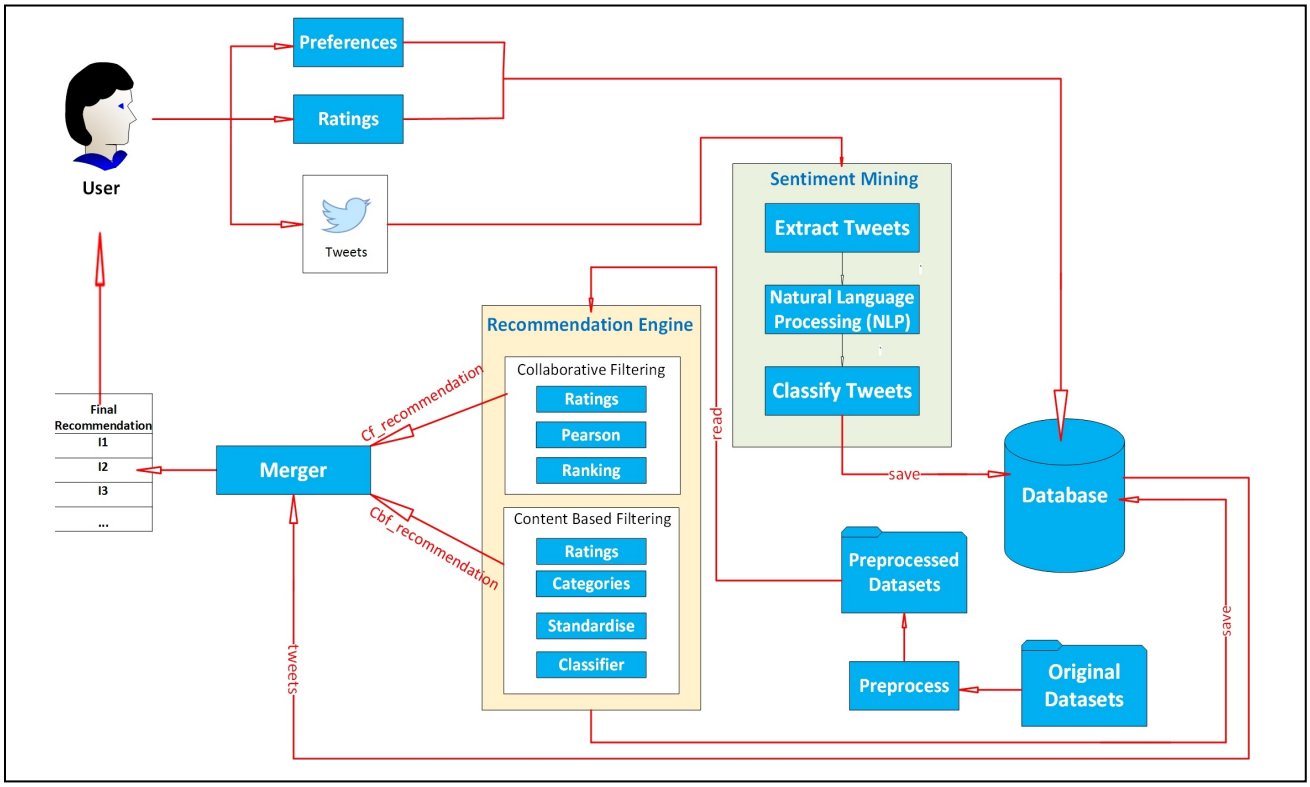


Fig. 1: System overview

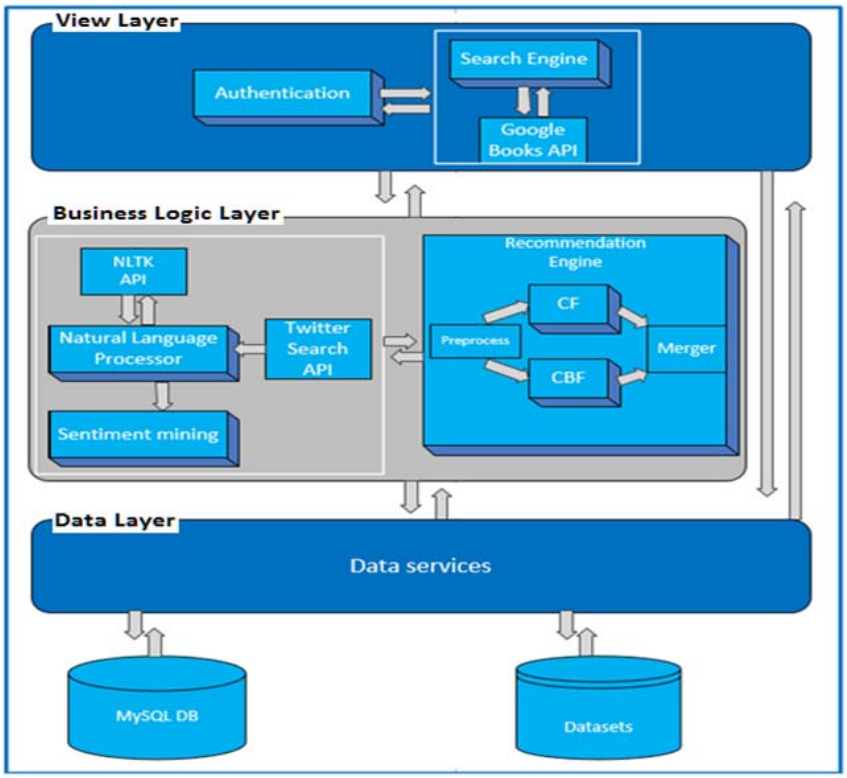


Fig. 2. System Architecture

F. *Merger*

The ultimate step in the recommendation engine is fusing the resulting recommendation lists generated from the CF and CBF module independently into a whole and combining preferences of the user from tweets using sentiment mining to provide a better recommendation list. The fusion is made by cycling through every single recommendation in the independent lists generated by both algorithms, identifying any potential similarity with each other. Any matching recommendation(s) is appended to the final list which is made available to the end user.

The algorithm for merger is as follows:

```
Begin
   Let CF  Rec be the recommendation of CF
   Let CBF  Rec be the recommendation of CBF
   Let tweet  cat be the categories of tweets of a user
   Compare CF  REC and CBF  REC
   Return common items as Rec
   Compare tweet  cat categories with Rec categories
   Return common items as OUTPUT
End
```

IV. SYSTEM PROTOTYPE

The system has been implemented using Python. The proposed recommendation technique is designed in Python and the resulting recommended lists of books for each user are stored on a MySQL database. To test the recommendation engine a user-friendly mobile web interface has been implemented. The *My Recommendations* interface as shown in Fig. 3 displays the recommendations generated for a specific user when the data goes through the recommendation engine. The *PreferredBooks* interface as shown in Fig. 4 displays a final list of books based on a user's twitter posts' and preferences when going through the merger module.

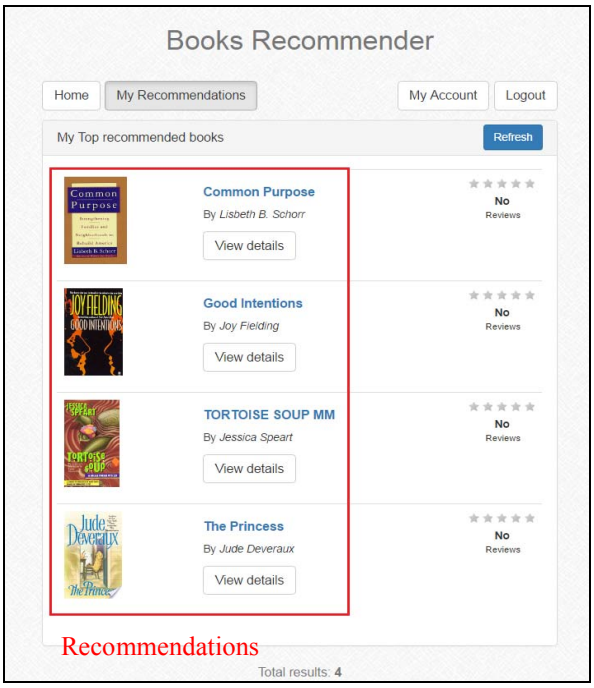


Fig. 3: User Recommendations

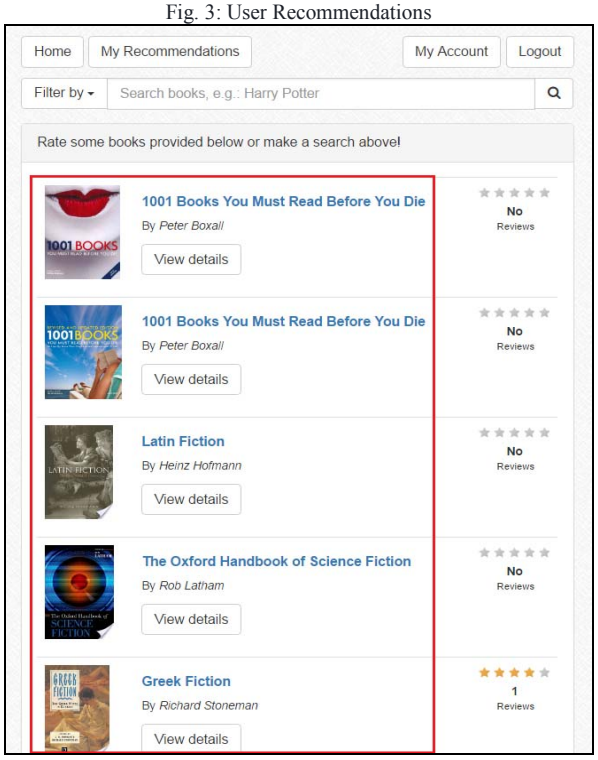


Fig. 4: Recommendations based on Twitter posts

Compared to the five systems described in section II, the proposed recommender system makes use of all the three recommendation techniques namely CF, CBF and sentiment analysis and captures data from the three data sources namely *User Profile*, *Open Source* and *Dataset*. It additionally takes into consideration aspects of Big Data. Both a desktop and a mobile version of the system have been implemented. The Scikit-learn model_selection [22] and classification_report [23] are used to obtain a classification report which includes certain useful metrics (Precision, Recall and Support) for evaluating the performance of the binary classification performed by the k Nearest Neighbors (kNN) Classifier. The report is based on a test set consisting of 813 records. Details of the classification report is summarized in Table 2. The classifier achieved a total precision of 67%. Out of the total 813 predictions made, 650 predictions are classified as correct (True) and 163 predictions are classified incorrect (False). The precision metric shows the exactness of the classifier [30]. In this paper, precision measures the correctly predicted books from the total predicted books. Higher values demonstrate high number of True positives, and lower values demonstrate lower numbers of False positives. The recall metric measures the percentage of correctly classified records [31]. It indicates a high completeness in deducing True predictions (0.96). Support is the percentage of data sequences that contain the pattern or the frequency that a record is present in the pattern [32]. The classification accuracy is 79.95% and has been calculated as follows:

Classification accuracy = [(No. of correctly predictions / total no. of predictions) * 100]
= [(650/813) * 100]
= ~80%

TABLE 2. CLASSIFICATION REPORT

	PRECISION	RECALL	SUPPORT
False	0.16	0.03	163
True	0.80	0.96	650
Average / Total	0.67	0.77	813

V. CONCLUSION & FUTURE WORKS

In this research work, a recommendation system has been implemented that makes use of CF and CBF together with sentiment analysis. The system extracts sentiments from Tweets using Twitter Search API and use ratings and preferences for items and apply data mining techniques to better understand customers. The system has been implemented in such a way that as soon as the tweets are added by users, they are being considered by the system to make proper recommendation. The performance of both the engine and sentiment mining module have been evaluated and observed that they perform within the right delay. Appropriate metric has been used to measure the performance of K-Nearest Neighbour classifier that is being used by content-based filtering. The results show that more accurate recommendations are being provided. Furthermore, CBF and CF and sentiment analysis from social media have been combined to provide relevant predictions. Currently the system supports a total of 153 categories of books. In future, the system will provide support for more categories. As the number of users increases in future, the data will increase in size so the system storage needs to cater for the growth of big data through the integration of Hadoop.

REFERENCES

[1] M. Chen, S. Mao and Y. Liu, "Big data: A survey", Mobile Networks and Applications, 19(2), pp.171-209, 2014.

[2] A. Ali, J. Qadir, R. ur Rasool, A. Sathiaseelan, A. Zwitter and J. Crowcroft, "Big data for development: applications and techniques", Big Data Analytics, 1(1), p.2, 2016.

[3] IBM Big Data & Analytics Hub. (2017). Why only one of the 5 Vs of big data really matters. [online] Available at: http://www.ibmbigdatahub.com/blog/why-only-one-5ĝ-vs-big-data-really-matters [Accessed: 07 August 2018].

[4] Z. Liao, Q. Yin, Y. Huang and L. Sheng, "Management and application of mobile big data", International Journal of Embedded Systems, 7(1), pp.63-70, 2014.

[5] G. NagarjunaReddy and M.J. Reddy, "Big Data Processing Using Hadoop in Retail Domain", International Journal Of Engineering And Computer Science(IJECS), pp. 18175-18179, 2016.

[6] D.F. Gurini, F. Gasparetti, A. Micarelli and G. Sansonetti, "A Sentiment-Based Approach to Twitter User Recommendation", In RSWeb@ RecSys, 2013.

[7] Laccir.org. 2017. "Quality Health Information Retrieval: Improving Semantic Recommender Systems with Friendsourcing > Problem Statement". [ONLINE] Available at: http://www.laccir.org/laccir/r1210lac007/ProblemStatement/tabid/800/Default.aspx. [Accessed: 07 August 2018].

[8] B. Valois Jr and M.A.D Oliveira, "Recommender systems in social networks", JISTEM-Journal of Information Systems and Technology Management, 8(3), pp.681-716, 2011.

[9] ReadWrite. 2017. 5 Problems of Recommender Systems - ReadWrite. [ONLINE] Available at: http://readwrite.com/2009/01/28/5_problems_of_recommender_systems/. [Accessed: 07 August 2018].

[10] J.P Verma, B. Patel and A. Patel, "Big data analysis: recommendation system with Hadoop framework", IEEE International Conference on Computational Intelligence & Communication Technology (CICT), (pp. 92-97). IEEE, 2015.

[11] K. Saranya and D. Sadasivam, "A Survey on Personalized Recommendation Techniques", International Journal on Recent and Innovation Trends in Computing and Communication, 2(6), pp.1385-1395, 2014.

[12] D. C. Anastasiu, E. Christakopoulou, S. Smith, M. Sharma and G. Karypis, "Big Data and Recommender Systems", 2016.

[13] M. Kohar and C. Rana, "Survey Paper on Recommendation System", International Journal of Computer Science and Information Technologies, 3(2), pp.3460-3462, 2012.

[14] H. Drachsler, H. Hummel and R. Koper, "Recommendations for learners are different: Applying memory-based recommender system techniques to lifelong learning", 2007.

[15] R. Lad, J. Lal, M. Laad, K. Thanvi, "ShopZoo: A Grocery Shopping Mobile Application with Smart Recommendation System", International Journal on Recent and Innovation Trends in Computing and Communication, vol. 4, pp. 48-51, 2016.

[16] J. Beel, S. Langer, B. Gipp and A. Nürnberger, "The architecture and datasets of Docear's Research paper recommender system", D-Lib Magazine, 20(11/12), 2014.

[17] N. Cao and W. Cui , "Overview of Text Visualization Techniques", In Introduction to Text Visualization (pp. 11-40), Atlantis Press, 2016.

[18] Y. Patel and V. Patel, "A Survey on Various Techniques of Recommendation System in Web Mining", IJEDR, 3(4), pp.695-700, 2015.

[19] M. Fleischman and E. Hovy, "Recommendations without user preferences: a natural language processing approach", In Proceedings of the 8th international conference on Intelligent user interfaces (pp. 242-244), ACM. 2003.

[20] A. Sachan and V. Richariya, "A survey on recommender systems based on collaborative filtering technique", International Journal of Innovations in Engineering and Technology (IJIET), 2(2), 2013.

[21] Hybrid Recommender Systems, Recommender Systems, 2016. [Online]. Available: http://recommender-systems.org/hybrid-recommender-systems/. [Accessed: 07 August 2018].

[22] Scikit-learn.org. (2017). sklearn.model_selection.GridSearchCV — scikit-learn 0.18.1 documentation. [online] Available at: http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html [Accessed: 07 August 2018].

[23] Scikit-learn.org. (2017). sklearn.metrics.classification_report — scikit-learn 0.18.1 documentation. [online] Available at: http://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html [Accessed: 07 August 2018].

[24] K. R Shrote and A.V Deorankar, "Hotel Recommendation System Using Hadoop Framework", Indian Journal of Applied Research, 6(4), 2016.

[25] R. Torres, "Combining collaborative and content-based filtering to recommend research papers", Porto Alegre, 2004.

[26] A. Felfernig, M. Jeran, G. Ninaus, F. Reinfrank, S. Reiterer and M. Stettinger, "Basic approaches in recommendation systems", In Recommendation Systems in Software Engineering (pp. 15-37). Springer Berlin Heidelberg., 2014.

[27] A. Agarwal, B. Xie, I. Vovsha, O. Rambow and R. Passonneau, "Sentiment analysis of twitter data", In Proceedings of the workshop on languages in social media (pp. 30-38), Association for Computational Linguistics, 2011.

[28] S. Vinodhini, V. Rajalakshmi and B. Govindarajulu, "Building personalised recommendation system with big data and hadoop MapReduce", Int. J. Eng. Res. Technol, 3(4), 2014.

[29] www2.informatik.uni-freiburg.de. (2017). Book-Crossing Dataset. [online] Available at: http://www2.informatik.uni-freiburg.de/~cziegler/BX/ [Accessed: 07 August 2018].

[30] L. Soibelman, J. Wu, C. Caldas, I. Brilakis and K.Y Lin, "Data analysis on complicated construction data sources: vision, research, and recent developments", In Intelligent Computing in Engineering and Architecture, Springer, Berlin, Heidelberg, pp. 637-652, 2006.

[31] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations", International Journal of Data Mining & Knowledge Management Process, 5(2), p.1, 2015.

[32] K. Rajput and B. Oza, "Comparative Study of Classification Techniques in Data Mining", International Journal of Creative Research Thoughts, 5(3), 2017.

Development of an IR-based Device for Wireless Communication in Community Health Centres

Chidi E. Ngene
Electrical and Electronics Engg. Science
University of Johannesburg
Johannesburg, South Africa
chidingene2279@gmail.com

Kayode E. Adetunji
Electrical and Electronics Engg. Technology
University of Johannesburg
Johannesburg, South Africa
kayvins@gmail.com

Thokozani Shongwe
Electrical and Electronics Engg. Technology
University of Johannesburg
Johannesburg, South Africa
tshongwe@uj.ac.za

***Abstract*—With the increase in implementation of advanced communication equipment, especially in the medical sector, finding solutions to anticipated problems has become paramount. Our envisaged problem is the uneasy procedure of taking patients' variables frequently in community health centers, where there is usually minimal support staff. This paper indicates that human variables such as body temperature and movement at a special ward in a community health care center are transferable from one computer to another. However, our implementation integrated a temperature sensor (LM35), pyroelectric infrared, and infrared to a low-cost Arduino-generic Lilypad microcontroller (as the transmitter). The developed device was able to communicate with the receiver side (comprised of Arduino Mega, infrared receiver, and a computer). We were able to transfer data seamlessly from Mr. C. E. Ngene to a remote computer for analysis. This project will make up for the unavailability of expensive devices and low workforce in the rural community health centers. This designed device can help patients to be attended to quickly as the patients start to have issues.**

Keywords—health, remote monitoring, infrared, optical wireless communication, sensors

I. INTRODUCTION

Data communication is the process of using certain techniques to transfer data (or information) from one point to another, in an electronic form. Information dissemination has been a crucial part of our everyday lives, especially in the electronic application for dynamic purposes. Over time, communicating data has evolved enormously, with the invention of copper wires, optical fibers, radio frequency (RF), infrared (IR), Wi-Fi, and GSM/GPRS. However, these types of communications have their challenges and have been researched effectively. The evolution of these technologies have seen wireless communication having more advantages. Optical communication has enhanced the capacity of wireless communication in terms of range. Optical wireless communication (OWC) is a method of optical communication that uses unguided rays or light (detectible, infrared, or ultraviolet) to carry signals from point to point. OWC is seen as an emerging technology that complements the radio frequency communication in some applications [1].

Wireless communication systems gain their efficiency using millimeter wave and sub-millimeter wave carrier frequencies to actualize a direct line of sight connections having a high speed and data rate of up to 100 Gbps currently [1]. We observe that over the years providing high speed, reliable data in Airplanes has been a major difficulty with the passengers on board considering Telecommunication as a communication network. A small fraction of network (Wi-Fi) connectivity is always received by a very few passengers at about 10-12% in the flight using an internet provider. This limits the Internet of Things (IoT) to some important areas inside the Airplanes, electromagnetic interference (EMI), and bandwidth availability (of 100Mbit/s) to all users on board, due to the fact that it is cost intensive [2]. Due to cost considerations, infrared radiation has supplanted other forms of wireless communication such as Bluetooth and Wi-Fi [3]. This technology, however, faces challenges such as the inability to achieve point-to-point communication due to in flock of people, even at the 1m distance thereby causing un-correctable errors [4]. Considering that IR rays are directed to IR diode point to point nodes positioning them using a microcontroller that is low cost to achieve its aim. A short wavelength makes it easier to be absorbed than to be reflected. Zeng et al [5] stated that IR communication system has a challenge of intersymbol interference (ISI) considering the presence of certain organism, thereby not finding its optimal direct communication. Related works is presented in Section II. Methodology is described in Section III. Results and discussions are presented in Section IV. Finally, the conclusion is given in Section V.

II. RELATED WORKS

Underwater wireless optical communication (UWOC) is a method of sending and receiving data under water using low data rates considering acoustic wave as a means of transmission [6 - 9]. UWOC integrates with visible light laser diodes (LDs) to demonstrate the interconnectivity of both technological applications using emission wavelengths in the blue to green spectral regime inside water regardless of the conditions. This will undermine the effects of free-flowing particles in turbid underwater. However, the surroundings of the system performance was never thoroughly checked. Haas et al [10] combined beam-steered infrared light communication (BS-ILC), visible light communication (VLC) and LIFI to create a communication link between 320THz in the 400-700nm range. Visible light communication is achieved by modulating the data on the LED luminaries to transmit data to its destination and then demodulating the data on arrival. Considering the existing infrastructure, VLC threshold, a range of color entry to a matrix of characters, can be lowered to allow for easy flow of data. This will, however, cause a slow data transmission rate. This advent has caused numerous studies for high data transmission rate in wireless systems [11]. It was noticed that

broadband wireless communication systems find its way in the highly complex digital baseband and fast and accurate signal processing. Improvement was reported on how difficult to develop power-efficient analog front ends (AFE) including the analog synchronization blocks [12 – 13]. Some major companies in the mobile sector are able to forecast the increase of data traffic in the mobile sector, thereby leading to bigger wireless capacity inventions. This has been applied in the device to device communication, using the networks of mobile data for fast proximity and accessible data centers. It was stated that the wireless monitoring device, an infrared camera, captures the thermal image of the foot soles of a patient or person when his feet rest on it. Using an LCD of 10.4-inch to showcase a thermal image in the wireless monitoring device.

The microcontroller is used to program a device used to identify and check the temperature of twelve fixed points of the right and left foot sole. The system monitors the temperature of the left and right foot soles through the wearable device, sending the information via Bluetooth and subsequently comparing the temperatures in the two feet [14]. Seeing that the process works, the pictures, information, and thermographs are being sent to the web server of hospital and doctors through the communication device. Using two circuit system to actualize its functions in the communication environment as a wireless means, one is used to display light, and the other is used to send emails using a microcontroller [15]. Both circuit systems are connected through the serial port of the device used to run the circuit.

III. Methodology

This section comprises the experimental phase, done in the Electrical and Electronics Engineering laboratory of the University of Johannesburg. The circuit function was tested using one of the authors Mr C.E. Ngene to emulate a patient in a hospital. In this paper we shall referring to Mr C. E Ngene as the "patient". Infrared LED diodes were used to create a communication link. At the transmitter circuit a 14m6286 LED, PIR- HC-5R501, BC547 NPN transistor, 220 and 230 ohms resistors, Arduino wearable Lilypad, lm35DZ were used to build the transmitter circuit. The receiver circuit used an IR 3mm LED phototransistor, Arduino mega 2560 to actualize this point to point communication using infrared. The LED 14m6286 from the transmitter circuit is responsible for sending and modulating the ASCII tables or characters out through the electromagnetic waves. Arduino Lilypad is used in this project to sense and record the body temperature, and through the use of a code, it transmits data through the 14m6286 LED. A USB plug cable is used to send data to the Arduino lilypad via a computer before wearing it around the arm. 220 and 230 ohms resistors are used to resist the flow of current in the circuit. A 3.7V battery was used to power the entire transmitter circuit. We provided the battery with a battery holder to protect it and to make sure that it is safe when connected to the entire circuit and strapped to the body of a patients. LM35 is a temperature sensor. It senses the temperature of the body of a patient when there is a discharge in the heat of the body of a patient. It sense the heat as a discharge thereby turning the intensity of the voltage of the component to read the temperature body of the patient. PIR-HC-5R501 is a motion sensor. It is used to detect the body movement of the patient because of its ability to evaluate the heat flow through thermal resistance by a conductor or the flow of electric current. The receiver receives a modulated signal through electromagnetic waves, then demodulates the signal received back to character to display it at the serial monitor in the computer.

Figure 1 shows the complete device to be used to measure the body temperature. The device worn around the arm of the patient, senses the body temperature through the heat emanating from the body of the patient, and the patient's movement. Afterwards, the signal will be sent to the receiver circuit, displaying the level of the patient's temperature and frequency of movement. The device is suitable for community health centres, where there are less equipment and workforce. The device works with an optimum radial distance coverage of 5.7 meters, spanning through a low-levelled glass door. With an infrared communication device having a long distance range, nurses and doctors may be able to take the readings from a patient and make proper diagnosis.

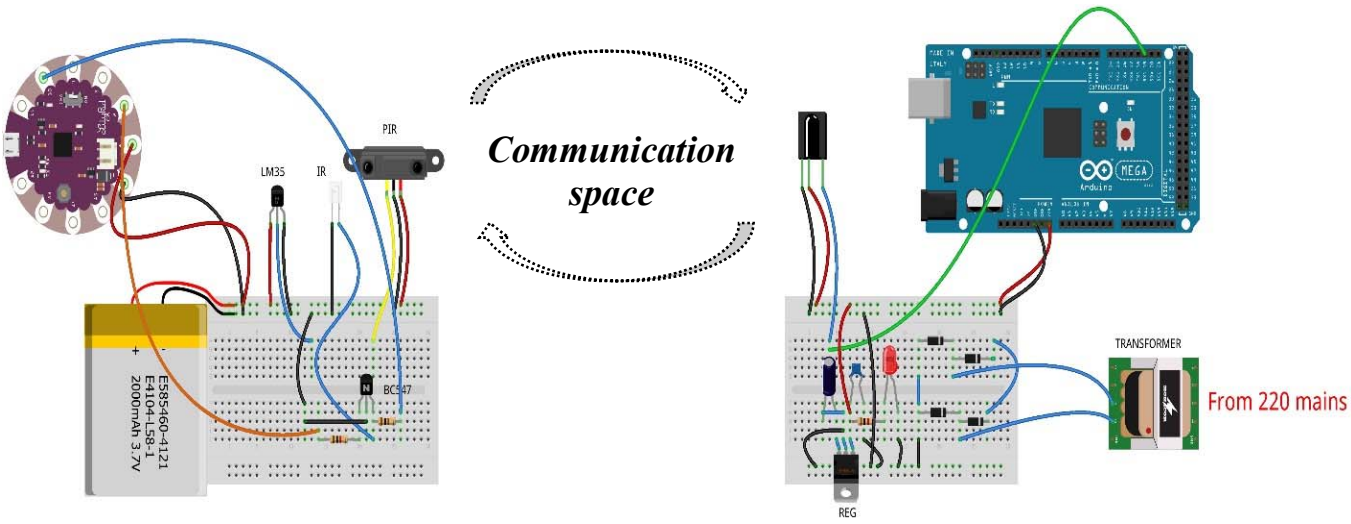


Figure 1: Developed work for implementation

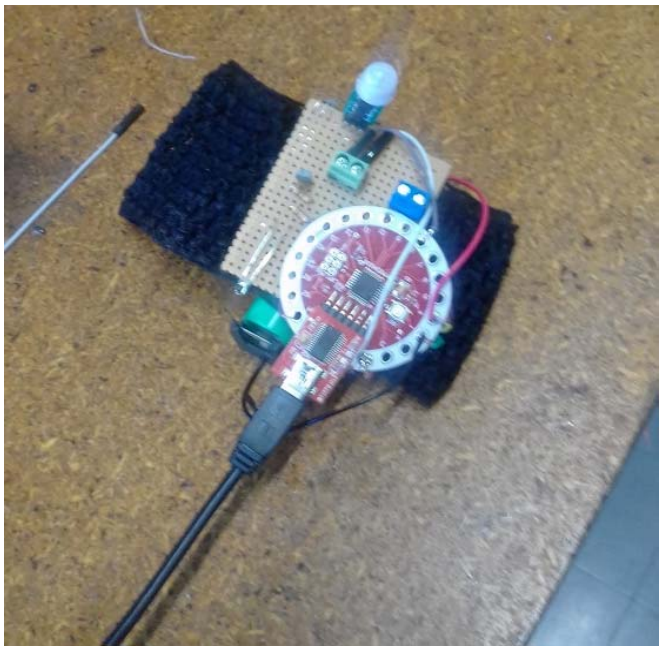


Figure 2: Arduino Lilypad circuit development

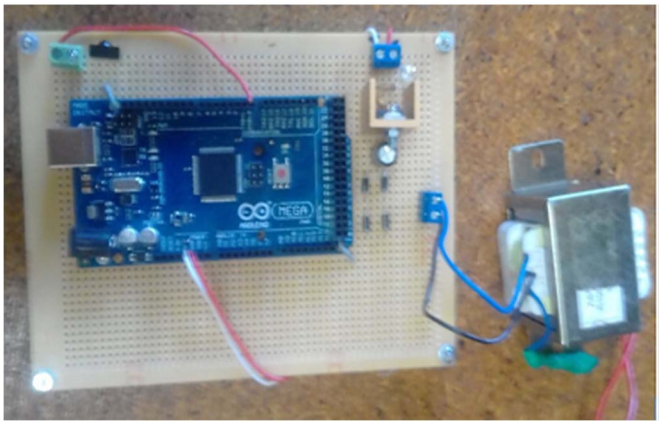


Figure 3: Receiver circuit of the project

IV. Results and Discussions

We discuss the results from the experiment and the implementation. The results below have been sectionalized into (a) performance evaluation and (b) testing.

A. Performance Evaluation

The prototype device was built and tested. It is observed that it can send and receive messages. A string "HelloWorld" was sent as ten characters as a packet, from the transmitter circuit to the receiver circuit. All ten characters were received back at a distance of .9 metres. The first character was received after 0.05 seconds, the second character was received after 05.47 seconds. It was observed that as the character was sent in packets from the transmitter circuit. The first character gained a momentum to reach the receiver circuit on time but the second and succeeding characters had a little delay in time to get to the receiver on time. Each received character were considered and evaluated. It was observed that the speed at which the characters reaches the receiver, decreases with time.

TABLE I. Character Sending

Characters	Number	Time (minutes:seconds:microseconds)	Speed (m/s)
H	1	00:00.05	1.8
E	2	00:05.47	0.16453
l	3	00:10.47	0.08596
l	4	00:15.23	0.0591
o	5	00:20.41	0.0441
w	6	00:25.19	0.03573
o	7	00:30.47	0.02954
r	8	00:35.37	0.02558
l	9	00:40.45	0.02225
d	10	00:45.19	0.01992

Table 1 above shows the characters, "HelloWorld" were sent as a packet, the number of characters are 10, the time each character was received was calculated and the distance used to send the characters was recorded which is 0.9m and the speed at which each character arrives was recorded as well. Figure 4 shows the graph obtained from the above experiments.

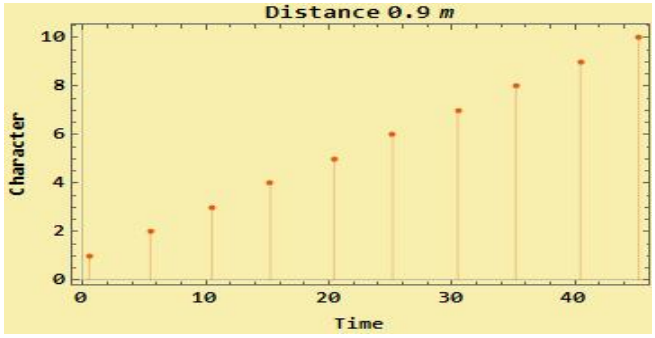


Figure 4: Character transfer against time

The graph above shows that characters were sent as a packet through the serial monitor from the transmitter and sequentially arrived at its destination. It was observed that the first character arrives at 0.05 seconds with a speed of 1.8m/s sub sequentially another character follows. The maximum optical power the first character arrives at the receiver increases as another character follows. It shows that there is a delay in the arrival of the character due to the baud rate used.

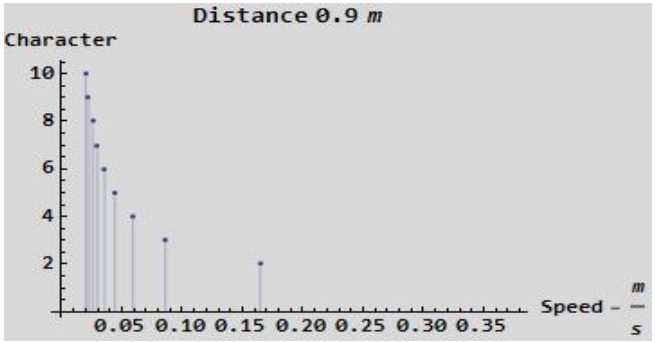


Figure 5: Character transfer against time

The graph above also shows that as the character is being transferred from the transmitter circuit to the receiver circuit at a distance of 0.9m, the characters tend to cluster to each other at the ending of the six characters. It is observed that the clustering of the character is as a result of the decrease in the speed of the arrival of the characters. It was noticed that there is a gap in the speed of the arrival of four characters and closeness in the arrival of the six characters. It shows that as the characters were sent in packets, optical heat used in the circuit first heat up the component used in sending the characters and slowly circulate to the other components, thereby reducing the speed at which the component send the characters.

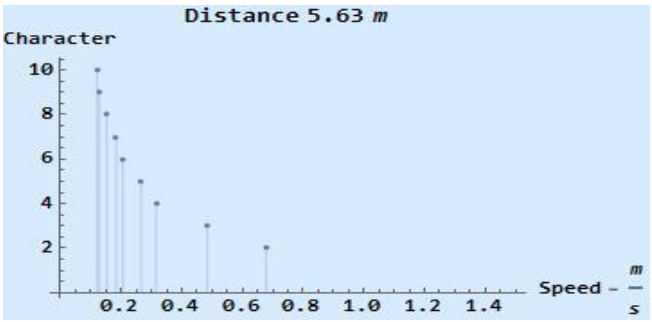


Figure 6: Character speed from transmitter to receiver

From the above graph in Figure 6, it is observed that there is a shift in the graph when the distance is 5.63m, being the maximum distance the receiver circuit will receive the character without any error. It is also observed that the characters sent in packets through the transmitter arrived with a speed of 2 seconds. However, the graphs from Figures 5 and 6 have a similar characteristics, even when there is a difference in time and distance.

B. Testing

The experiment was tested in the electrical laboratory at the University of Johannesburg, which confirmed the working of the device. The device was able to send and receive data. The testing phase showed that temperature and body movement of a patient in a health center can be remotely sensed for display in a computer through the receiver circuit. Figure 7 shows the device being strapped to a subject's arm for measurements.

Figure 7: Testing the device

TABLE II. RESULTS FROM THE DEVICE WORN TO THE PATIENT

Time (h:m:s)	Body Temperature (°C)	Body movement
16:21:37	36.8	0
16:22:37	36.7	0
16:23:36	36.7	1
16:24:37	36.7	1
16:25:39	36.5	1
16:26:37	36.7	0
16:27:38	36.9	1
16:28:39	36.9	1
16:29:39	36.9	0

°C – degrees Celsius.

Table II is a snapshot of readings from a patient, displaying the body temperature and movement, and the time received. Body movement measurements are binary, where "1" stands for body movement, and "0" stands for an approximately still body. Figure 8 shows the graph illustrating a snapshot of temperature readings from the patient.

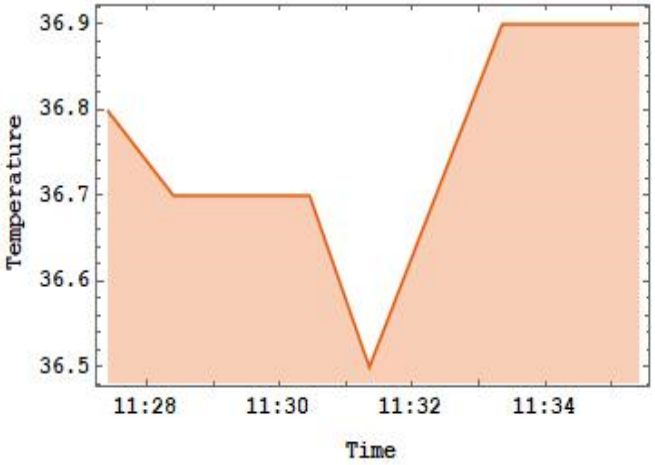


Figure 8: Temperature readings from patient

V. CONCLUSION

Optical wireless communication using infrared beams implemented in this paper support a wavelength radiation of 90º and a distance of up to 5.6m. We notice that infrared offers a huge un-shared capacity to independent point-to-point devices, with maximum power efficiency making good use of a microcomputer serial ports for automatic encoding and decoding functions. The use of a simple hardware circuit was used to achieve our main objective, which entails a low cost, uncomplicated software design, and stable performance of the system.

We suggest that for high-speed data transfer, the baud rate can be increased. Also, components can be replaced or added for higher reception or faster reception from the system setup. It is suggested that optical wireless communication using infrared systems eliminates the congestion in the radio spectrum by helping to offload high capacity traffic loads, thereby leading to a more cost-effective solution.

We realize that our system functions properly in sending and receiving data, especially very useful at community health centers with fewer equipment or facilities. We used signal modulating and demodulating with Arduino Lilypad to disseminate information to a computer for data display, using a baud rate of 1200bps to communicate a better data transfer. The future scope of this paper can be found, such as to measure high and low blood pressure.

REFERENCES

[1] S. Koenig, D. Lopez-Diaz, J. Antes, F. Boes, R. Henneberger, A. Leuther, A. Tessmann, R. Schmogrow, D. Hillerkuss and R. Palmer, "Wireless sub-THz communication system with high data rate," Nature Photonics, vol. 7, pp. 977, 2013.

[2] Gogo Inc., "The leading provider of inflight entertainment," vol. 2017, Feb 15. 2017.

[3] J.B. Carruthers, "Wireless infrared communications," Wiley Encyclopedia of Telecommunications, 2003.

[4] W. Chunyue, W. Lang, C. Xuefen, L. Shuangxing, S. Wenxiao and D. Jing, "The research of indoor positioning based on visible light communication," China Communications, vol. 12, pp. 85-92, 2015.

[5] Zhaoquan Zeng, Shu Fu, Huihui Zhang, Yuhan Dong and J. Cheng, "A Survey of Underwater Optical Wireless Communications," Comst, vol. 19, pp. 204-238, 2017.

[6] H.M. Oubei, C. Li, K. Park, T.K. Ng, M. Alouini and B.S. Ooi, "2.3 Gbit/s underwater wireless optical communications using directly modulated 520 nm laser diode," Optics Express, vol. 23, pp. 20743-20748, 2015.

[7] C. Shen, Y. Guo, H.M. Oubei, T.K. Ng, G. Liu, K. Park, K. Ho, M. Alouini and B.S. Ooi, "20-meter underwater wireless optical communication link with 1.5 Gbps data rate," Optics Express, vol. 24, pp. 25502-25509, 2016.

[8] H. Lu, C. Li, H. Lin, W. Tsai, C. Chu, B. Chen and C. Wu, "An 8 m/9.6 Gbps underwater wireless optical communication system," Signal, vol. 600, pp. 7GHz, 2016.

[9] K. Nakamura, I. Mizukoshi and M. Hanawa, "Optical wireless transmission of 405 nm, 1.45 Gbit/s optical IM/DD-OFDM signals through a 4.8 m underwater channel," Optics Express, vol. 23, pp. 1558-1566, 2015.

[10] N. Neumann, T.B. Keuter, M. Laabs and D. Plettemeier, "Carrier recovery for sub-millimeterwave wireless transmission," in Radio and Wireless Symposium (RWS), 2017 IEEE, pp. 90-93, 2017.

[11] C.V.N. Index, "Cisco visual networking index: Global mobile data traffic forecast update, 2015-2020 white paper," Accessed Date, 2016.

[12] S. Huang, Y. Yeh, H. Wang, P. Chen and J. Lee, "W-band BPSK and QPSK transceivers with Costas-loop carrier recovery in 65-nm CMOS technology," IEEE J.Solid State Circuits, vol. 46, pp. 3033, 2011.

[13] H. Haas, L. Yin, Y. Wang and C. Chen, "What is lifi?" J.Lightwave Technol., vol. 34, pp. 1533-1544, 2016.

[14] T. Amartuvshin, D. Erdenechimeg, N. Chuluunbaatar, T. Enkhbaatar and D. Enkhzul, "Preprocessing techniques for daily healthcare device using SHIMMER™ wireless sensor platform," in Strategic Technology (IFOST), 2016 11th International Forum on, pp. 558-562, 2016.

[15] Raspberry Pi, "Raspberry Pi 3 Model B - Raspberry Pi," vol. 2017.

Design of a Cloud-based Monitoring System for Potential Leaves Decomposition for Manures

Kayode E. Adetunji
Electrical and Electronics Engineering Technology
University of Johannesburg
Johannesburg, South Africa
kayvins@gmail.com

Chidi E. Ngene
Electrical and Electronics Engineering Science
University of Johannesburg
Johannesburg, South Africa
chidingene2279@gmail.com

***Abstract*— The agriculture sector is important to the economic growth of agriculture-inclined countries like South Africa. Issues regarding agriculture may not only directly affect consumers, but cause the increase in food price. This may further affect the price control policy on food. This paper discusses soil manuring as one of the problems faced in optimum food productivity. Since leaves droppings could be decomposed to fertilize the soil, the need to understand the factors that may affect the process or outcome is important, hence our project. We designed a compact monitoring circuit using the Redboard and GSM/GPRS module. Three TCS3200 color sensors were used to detect leaves droppings, giving a large coverage area. The colors detected were categorized to numbers, for possible data analysis. Other sensors were used to collect parameters for possible factors that may affect decomposition. We measured soil moisture, soil temperature, ambient temperature, relative humidity, and dew point. Our Heroku-deployed developed cloud platform was synced to the monitoring circuit for remote monitoring. Our whole project embraced a continuous real-time monitoring from the site to the cloud platform.**

Keywords*— *agriculture, remote monitoring, cloud, soil fertility, communication, sensors

I. INTRODUCTION

Food is an essential part of life, both to the high- and low-income population. This has enhanced the agricultural sector research and development to seek optimum ways to increase farm produce. Challenging factors such as climate change and erosion has been majorly featured in the quest to optimize productivity. It is therefore important to monitor closely these challenges [1]. It is imperative that the study of challenges in the agriculturally sector helps to reduce loss of crops and to understand when and how to apply chemical or fertilizer to help with the growth of the crops [2]. With the monitoring of these crops, it is observed that they tends to produce healthy leaves and crops from the soil. Taking into consideration that there are other factors that will affect productivity of crops such as, temperature, humidity. Monitoring the growth of crops and taking into accounts of the effects that helps in emancipation of some insect infected crops and to prevent them from being affected by using remote sensors to checkmate them [4]. Using remote sensors helps to combat the unnecessary effects on crops and improves the study of remote sensors by researchers.

One aspect of agriculture is the fertility of crops. This has come in many ways, being that soil fertility can be enhanced through chemically made fertilizers or natural fertilizers such as leaves and animal dungs. The former proves to be expensive for farmers in small scale farming. Sometimes, farmers use the in-situ method of decomposing green leaves [6]. Here, they bury leaves in the soil prior to crop cultivation. We are concerned about the latter, particularly because of its random availability. Decomposition of leaves' droppings can be beneficial to the soil through [3] (i) drought management to crops (ii) thickening of soils (iii) decreasing possible toxicity to crops through secretion of phosphorus, and (iv) reducing the effect of excessive acidity and alkanity.

Therefore, the need to research on the decomposition properties is essential. With the advent of research, remote control encounters problems when focusing on monitoring of network operational activities through a device, especially with the exposure of images to sunlight and its enlargement, recovering the message that happened in a far place through applying image as a pictures and optical information being one of the aim of remote monitoring [5]. It is observed that the closer objects are detected better than father objects. The lightning can be controlled with less complications, thereby giving us more features in the output of any object of concern that is channel towards monitoring facility. While distant object tends to appear smaller in size when viewed from a monitoring point, thereby giving a relatively low intensity of the constituent components while the closer item will give the full amount of its intensities [7]. Sequel to this observation, are the challenges faced in handling focus, disclosure, and image magnification for distant objects. There are efforts in research on how to produce a new device for a remote digital monitoring system to solve the problem of recording visual data [8].

Cloud computing is defined as the sharing of resources over the internet. Users take advantage of the resources shared in their use of the internet. In the resources shared, amount of work is shifted to reduce workload on traditional computers. In this technology, customers can choose a third-party provider for cloud services as an alternative than to set up their individual physical infrastructure. This results to the user paying only for the resources used for their service. The cloud is made up of networks of computers [9], which make it easier for the local computers because the load is not on them, hence reducing the handy network system on the user's side. Including some qualities of resource sharing that is on-demand for self-service in cloud computing, pricing, quality of service and resources pooling.

We look at three levels in cloud usage namely, (i) using the third party as a provider, (ii) providing a platform that allows customers to manage some applications and running the service on a system without installation of software on the

local computer. Generally, many customers actually incorporated cloud computing in the form of services, using Gmail, Twitter, Yahoo and Hotmail. This improves production and innovation, reducing the responsibility of reduction in management and meeting the needs of changing the environment as their businesses change. Defining cloud computing as the offspring of dispersed, shared resources for others to benefit from through networking by some authors in the research of cloud computing [9].

II. Remote monitoring mechanism

Remote monitoring (RM) is a way of monitoring a particular system without having to be physically present. It can be termed as observing a system for different purposes. It is important to understand a system and not having to be at the physical site, which can differ in distance. The distance is always dependent on the type of technology used (such as Bluetooth, Infrared, Radio frequency). However, the cloud-based monitoring system does not consider distance since communication can be reached anywhere there is an internet connection. This type of communication (cloud-based) is mostly carried out using Wi-Fi or GSM/GPRS technology.

The main component of this mechanism is the data logger or a data acquisition system. The data loggers sometimes called data acquisition systems (DAC) are portable computing devices that receive and store data from a source and at a pre-programmed interval. Data loggers have two main parts, which are the microcontroller board and the storage unit. The external memory stick or SD card form the part of the storage, and may be used to load a program to the flash memory of the DACs' microprocessor. A web-based data logger uses Internet protocol (IP) and ports for communication and location, it has a private and public IP address for the allowance of the internet. Data loggers have been used for sensing and monitoring [10] [11] [12].

Cloud-based RM platforms are generally divided into sections, which are data collection and logging, data processing and set-off, and storage and access section. Data collection and logging deals with the acquisition of data (such as temperature, humidity, pressure, voltage, and current) from applicable sensing devices. This section may comprise of PV panels, meteorological instruments, sensors and data loggers. Data processing and setoff section deals with the processing of the raw data into a file format for a readable view. It also acts as a buffer with a predefined interval and afterward transfers the processed data to a server. Finally, the storage and access section involve cloud server, database, and web application (for viewing data). This section may be called the data monitoring or virtualization platforms, and they are used for accessing data that has been sent from a pre-developed web application. The platform enables viewers to visualize data graphically or in tabular form, using Application Programmable Interface (APIs) to structure architecture and make them vary from one another. These APIs can be written as HTML/CSS/JavaScript, C++, Python, or Scala.

III. Remote Monitoring in Agriculture

There have been proposed and developed remote monitoring platforms to either explore or solve a problem. These platforms have their type of configuration, which may depend on the location, size of project, or cost effectiveness. The problem to crack is also an important factor to the type of platform developed.

Nooriman et al [13] used wireless monitoring network (WSN) to monitor, control, and manage a specific specie of plant (*Harumanis* Mango). They measured soil moisture, temperature, and humidity by utilizing a PIC18 microcontroller and an RF transceiver module. The experiment was done to understand the environmental conditions that may affect a type of crop. Similarly, Singh et al [14] measured soil moisture and rainfall to determine the growth of crops. It was envisaged that soil moisture plays an important role in the healthy yield of crops. They used ZigBee Development kit (by Texas instrument) for measurement of weather variables, and an LCD screen for quick monitoring. However, the viewing of reading on-site may be restricted to farmers only. The system may also pose difficulty for farmers that live far from the farm. Another work [15] proposed the design of a vegetable monitoring system for greenhouse farming. They also made use of the ZigBee technology but incorporated GPRS communication for real monitoring. This was a step further compared to the previous works. The web application was developed using ASP.NET and modular components for dynamic page display. The system showed a reliable error value, when measurements were compared to the physical monitoring approach.

Recently, Adetunji and Joseph [16] developed an RM platform to observe readings from an agricultural resource. Soil moisture, soil temperature, ambient temperature, relative humidity, and dew point, were measured and stored remotely. The aim was to develop a system that will have potential foundation for exploration of data analysis, especially in the era of big data. The platform was developed using Node.js server runtime for the backend server, reactjs for frontend application. They used the HTML library, chartjs to visualize data. They laid emphasis on understanding the nature of the soil for healthy crops. Other works in agriculture are based on irrigation. Authors [17] [18] have implemented automatic irrigation using environmental parameters through sensors. Pushkar and Sanghamitra [1] proposed a relating factor in the research of agriculture and its importance. They used Wi-Fi module to send and receive information to a browser. Integrating Arduino to enhance the performance of water system for smart irrigation with a simple sensor for soil moisture. Two microcontroller device were used to monitor the water content of the soil. These microcontroller shows when a dry and wet soil is in good or bad conditions.

IV. Objectives

The work aims towards monitoring leaves' droppings for decomposition to manure. It has been envisaged that leaves falling off from trees or plants can be decomposed to form nutrients for the soil [4]. This decomposition can be hastened by the air temperature, humidity, soil moisture, and soil temperature. Hypothetically, leaves, animal dungs, and other

kinds of manure-degradant, will start to process into manuring properties only when they match the color of the soil. The monitoring and analysis of environmental parameters can help understand more about decomposition practices.

The objectives of this paper are to:

1. Develop a model to practicalize the measurement of leaves colors compared to related factors
2. Design an on-site sensing circuit to measure soil moisture, soil temperature, ambient temperature, relative humidity, and dew point
3. Develop a cloud platform to hold a backend application (and a database)
4. Create a frontend application to analyze the received data from the site.

V. Methodology

This work is based on development and implementation of a monitoring platform for agricultural use. In this section, we discuss the materials, methods, and type of deployment used to efficiently achieve the aim of this paper. The design of the project was based on (i) portability of the device, (ii) remote sensing, and (iii) accuracy of measurements.

We deployed our developed circuit on a farmland, where there are shrubs and frequent leaves droppings from plants and trees. Figure 1 shows the land area in Nkangala district, Mpumalanga Province.

Figure 1: Example of a site in Mpumalanga with leaves droppings

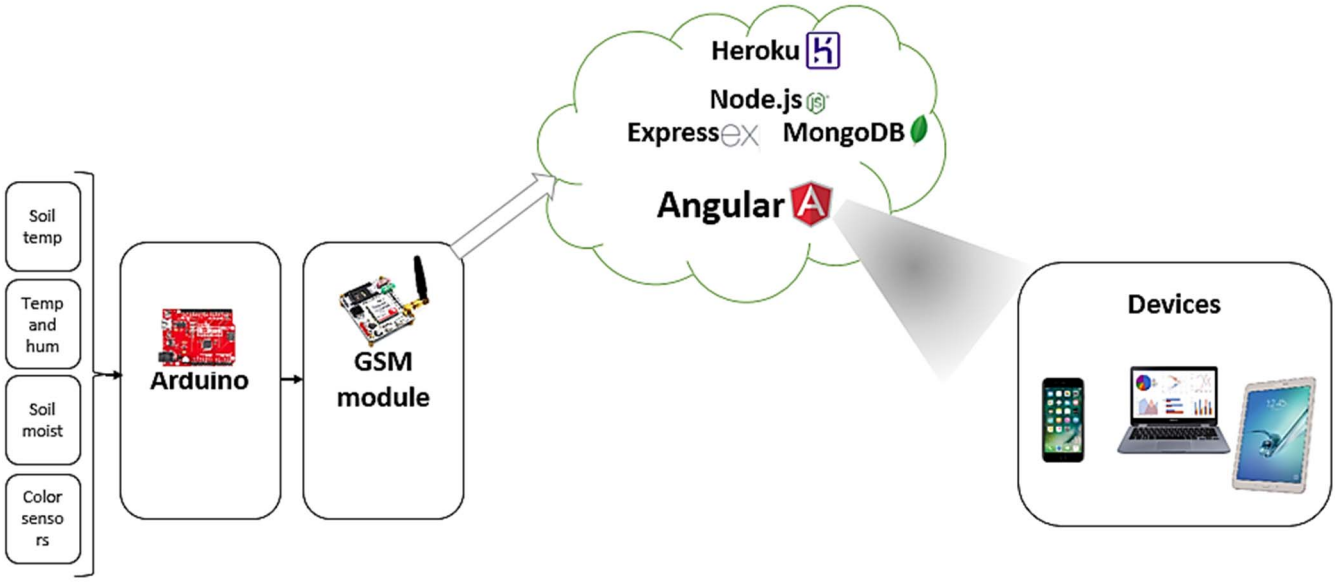


Figure 2: Overall project map-out

The GSM module was used to reduce hardware. Wi-Fi modules will always require a router, which is an additional hardware. All sensors were connected to the Arduino for parsing data. The data were simultaneously parsed through the Transmit (Tx)/Receive (Rx) pins. The sending of data from the GSM module to the cloud platform, ends the function of the on-site platform. Other sections discuss the electronic design and the cloud implementation.

A. Electronic Circuit Design

This section discusses the schematic layout of the project. The Arduino Mega was used as the microcontroller, which was the primary basis for data acquisition. The medium for communication between the on-site deployment and cloud platform was the GSM module. Figure 1 shows the schematic diagram for the connections.

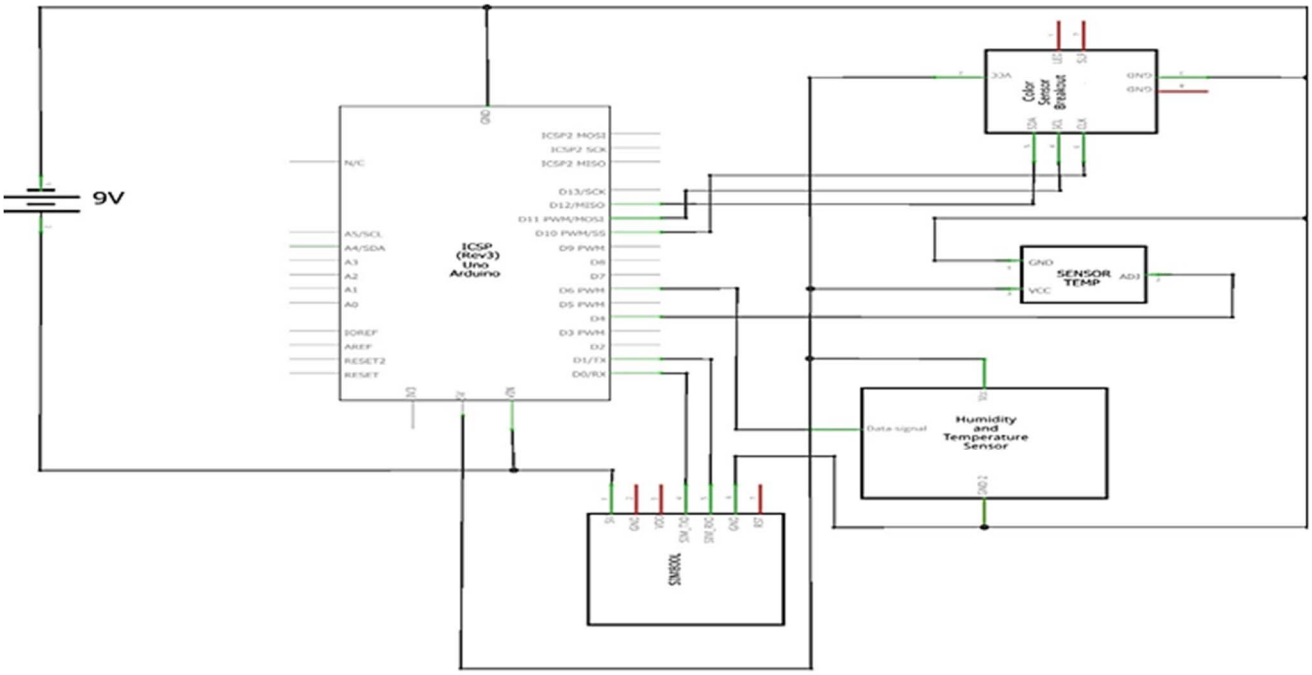


Figure 3: Schematic Diagram for the on-site development

The assembly involved connection between the Redboard and sensors such as DHT22, DS18S20, and TCS3200. The circuit from Figure 3, was designed with the Fritzing software using the schematic view approach. The Arduino and GSM module were powered through the 9 volts battery. All sensors are powered through the 5 volts output from the Arduino. The DHT22 temperature and humidity sensors senses the temperature, humidity, and dew point of the environment. DS18S20 measures the temperature of the soil, while the TCS3200 was used to sense colour of leaves droppings.

The TCS3200 is an 8-pin color sensor, which uses photodiodes to detect colors. Primary and secondary colors are being detected through the S2 and S3 pins being photodiodes to be displayed in the computer. The S0 and S1 pins are used for the frequency scaling of the sensor. The frequency is scaled in 2%, 20% and 100% according to the pin of the TCS3200 color sensor. The color sensor detects light color through the frequency, using the intensity of light to detect different colors. The colors are being read by the use of the photodiodes present in the TCS3200. It is compatible and can be programmed with most microcontrollers. Snippet 1 shows the implementation of the TCS3200 color sensor.

```
1.  void setup() {
2.  digitalWrite(S0, HIGH);
3.  digitalWrite(S1, LOW);
4.  }
5.  void loop() {
6.  digitalWrite(S2, HIGH);
7.  digitalWrite(S3, HIGH);
8.  frequency = pulseIn(sensorOut, LOW);
9.  green = map(frequency, 100,199,255,0);
10. if (green > yellow && green > brown){
11. GSM.print(green);
12. }
```

Snippet 1: Snapshot of color sensor code

Snippet 1 illustrates the method of receiving green color from the color sensor. The `pulseIn()` function is used to read the frequency. Afterwards, `map()` was used to convert the frequency readings to values between 0 and 255. The flowchart of the project is displayed in Figure 4.

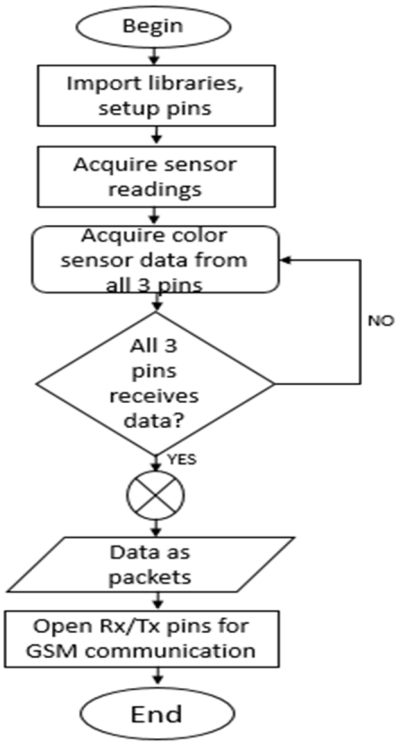


Figure 4: Flowchart of the on-site development

B. Cloud Implementation

This section discusses most of the abstract part of the project. The deployment of the type of server runtime, web application, and database were discussed.

1) Backend Application

Node.js was as the server runtime. This software uses an asynchronous approach to send and receive server messages (in the form of request and response). This approach makes traffic denser especially at peak periods. Unlike the traditional servers, that is, synchronous, multithreaded servers (e.g. Apache), requests are processed sequentially by independent units. The units have specific tasks, which enables requests to be executed by "*departments*". This action makes it require less hardware for processing units (PUs). The server runtime was developed with an Express web application framework. This enables us build web servers for listening to incoming requests, and solving them for appropriate response. Routing is another basic function, which makes it possible for switching views on a webpage.

MongoDB database was deployed for storing data. MongoDB is a NoSQL database that allows heavy traffic of data with minimal crashes. Hsu et al [19] had defined NoSQL as a database that is built to provision heavy read and write tasks. They are fit to run in cloud environments because of their capabilities. MongoDB has a module called Mongoose, which helps manage relationships between data. MongoDB was installed using the npm install from Node.js. The Mongoose was also installed using `npm install mongoose -save`. Promises were used in the backend application (Express app) to link our database, and our database schema is defined as shown in Snippet 2.

```
let dataSchema = new mongoose.Schema({
    soil_moisture: number,
    soil_temp: number,
    air_temp: number,
    humidity: number,
    leaves_color: String
});
```

Snippet 2: Data schema creation

The leaves_color from the array was converted to a number format, using

```
if (leaves_color == "green") {
   let leaf = parseInt(leaves_color, 1);
}
else if (leaves_color == "brown"){
   let leaf = parseInt(leaves_color, 2);
}
else if (leaves_color == "yellow"){
   let leaf = parseInt(leaves_color, 3);
}
else {};
```

Snippet 3: Snapshot of leaf color categorization

This is to create a platform for data analysis, as strings does not conform to certain analytical conventions. The schema was built on a model using the syntax:

```
let data = mongoose.model("data", dataSchema);
```

Other tasks were done from the Express App using the HTTP GET request.

2) Frontend Application

Googles' Angular framework technology was implemented. Angular is based on Typescript programming language, which passes functions like decorators, constructors, and static typing. Like other JavaScript frameworks, Angular is composed of components that makes web development segmentalized. Angular offers a two-way data binding with its Model-View-Controller (MVC) architecture. This enables a predefined model to controlled by view, which means that data delivery will change the model and vice-versa, according to the pre-programmed code. This model easily works with virtualization modules like D3.js. While Angular can be connected to the backend app, it could also directly interact with the database.

3) Virtualization

D3.js was used for visualizing data from the database. It is an open source JavaScript library that helps to create visual representation of data. Codes are embedded in HTML scripts tags and CSS. We have integrated D3.js with Angular app to bind data from Angular. After installing D3js, `with npm install d3`, it was imported to the app.component.html before using it in the HTML element.

VI. Results and Discussion

We discuss the final layout of project and the results from our deployed cloud platform. Figure 5 shows the site where the project is implemented.

Figure 5: Site of the developed project

Figure 6: Circuitry box and the TCS3200 color sensor

In Figure 5, the color sensors (in white) were strapped to trees. These sensors were separately connected to the Arduino, together with other sensors (soil moisture, soil temperature sensor, air temperature, and humidity sensor). The whole circuitry was boxed in a casing, with punched holes for sensor wires to connect through. Other color sensors (2) were connected to the circuitry box through wires beneath the soil. Figure 7 displays data points from the site.

Choose File

Points	Humidity	Dewpoint	Air temperature	Soil temperature	Soil moisture	Leaves Color	Timestamp
46	64.8	5.35	12	10.44	92.69	1	2018-06-11 21:00:21 UTC
47	63.8	4.93	11.8	10.5	92.45	1	2018-06-11 21:30:43 UTC
48	65.2	5.35	11.9	10.5	92.22	1	2018-06-11 22:00:33 UTC
49	64.3	4.85	11.6	10.5	91.32	1	2018-06-11 22:30:25 UTC
50	64.3	4.95	11.7	10.5	91.07	1	2018-06-11 23:38:25 UTC
51	65.3	5.08	11.6	10.5	91.06	1	2018-06-11 23:31:25 UTC
52	65.4	5.01	11.5	10.5	91.46	3	2018-06-11 00:00:24 UTC

Figure 7: Webpage displaying data points

From Figure 6, it is seen that all parameters sent from the site is displayed. Data can be downloaded as csv files. The graph is displayed in the Figure 8 and 9 shows the graphs from the frontend.

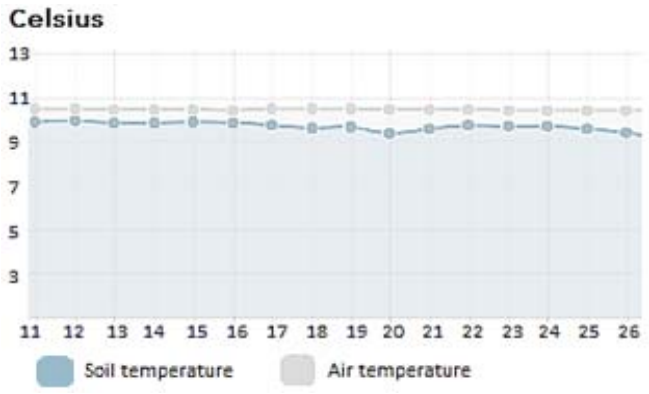


Figure 8: Simultaneous Readings from soil and air temperature

Figure 8 shows the simultaneous readings from soil and air temperatures. It is seen that the soil temperature is continually below the air temperature. This might not be entirely factual, as it may be reliant on the seasons. From [16], soil temperature was higher than the air temperature caused by the summer season.

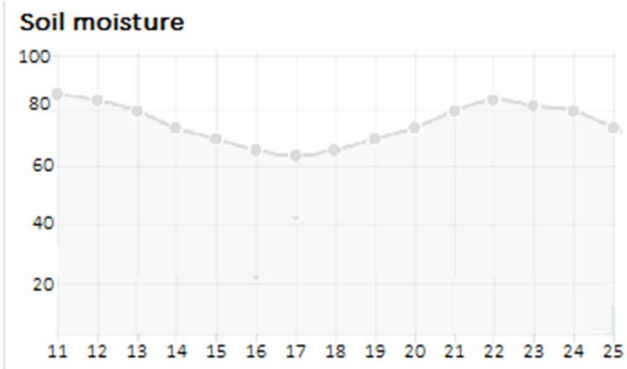


Figure 9: Soil moisture readings

From Figures 8 and 9, show a snapshot of the continuous real-time update from the Angular application. The x-axes are data points, or the entry id of all measurements. The graph could give more insight about the relationship between both variables.

VII. Conclusion

The abundance of leaves droppings will continually increase, which may pose a waste level. However, understanding the decomposition factors of these leaves may create a possibility to increase the adoption of these leaves' droppings for manuring. The work presented in this paper illustrates the successful implementation of cloud-based monitoring platform, which uses a minimal GPRS connection to measure and send environmental parameters to the cloud. Three TCS3200 color sensors were used to cover a wide area of farmland, and summed up in our algorithm to produce a near accurate color detection. The use of color sensors to detect color of leaves was cost-effective compared to installing cameras for vision. Recommendation for future works is to analyze the change in leaves color with respect to factors such as soil moisture, ambient temperature, soil temperature, humidity, and dew point.

References

[1] P. Singh and S. Saikia, "Arduino-based smart irrigation using water flow sensor, soil moisture sensor, temperature sensor and ESP8266 WiFi module," in Humanitarian Technology Conference (R10-HTC), 2016 IEEE Region 10, pp. 1-4, 2016.

[2] S. Dey, N. Karmakar, R. Bhattacharyya and S. Sarma, "Electromagnetic characterization of soil moisture and salinity for UHF RFID applications in precision agriculture," in Microwave Conference (EuMC), 2016 46th European, pp. 616-619, 2016.

[3] Z Z. Qingbo, "Status and tendency for development in remote sensing of agriculture situation [J]," Journal of China Agricultural Resources and Regional Planning, vol. 5, pp. 002, 2004.

[4] A. Vakeesan, T. Nishanthan and G. Mikunthan, "Green manures: nature's gift to improve soil fertility," Leisa-Leusden-, vol. 24, pp. 16, 2008.

[5] D. Riyana, Y. Widiyastuti, H. Widodo and E. Purwanto, "Effect of manure and plants spacing on yield and flavonoid content of Elephantopus scaber L." in IOP Conference Series: Earth and Environmental Science, pp. 012038, 2018.

[6] Li, M.; Lavest, J.M. Some Aspects of Zoom Lens Camera Calibration. Technical Report ISRN KTH/NA/P—95/03—SE, Dept of Numerical Analysys and Computing Science, KTH (Royal Institute of Technology, February 1995.

[7] J. Orlosky, T. Toyama, K. Kiyokawa and D. Sonntag, "Modular: Eye-controlled vision augmentations for head mounted displays," IEEE Transactions on Visualization & Computer Graphics, pp. 1, 2015.

[8] Y. Yao, B.R. Abidi and M.A. Abidi, "Extreme Zoom Surveillance: System Design and Image Restoration." Journal of Multimedia, vol. 2, pp. 20-31, 2007.

[9] M. Ali, S.U. Khan and A.V. Vasilakos, "Security in cloud computing: Opportunities and challenges," Inf.Sci., vol. 305, pp. 357-383, 2015.

[10] N. Patrascoiu and I.C. Barbu, "Data logger Built using the MS8050 bench multimeter and virtual instrumentation," in Carpathian Control Conference (ICCC), 2016 17th International, pp. 836-841, 2016.

[11] S. Mazhar, J. Gul, F. Mueen, and M. Hussain, "Design of a memory-card based low-cost GPS data-logger for livestock monitoring," in SENSORS, 2015 IEEE, pp. 1-4, 2015.

[12] S. Shashidhar, K.R. Sourabha, S. Santhosh, V. Nafeez and M.P. Sunil, "Design and fabrication of a low-cost single board portable GPS data logger," in Electrical, Computer and Communication Technologies (ICECCT), 2015 IEEE International Conference on, pp. 1-6, 2015.

[13] W. M. Nooriman, A. H. Abdullah, N. A. Rahim and K. Kamarudin, "Development of wireless sensor network for Harumanis Mango orchard's temperature, humidity and soil moisture monitoring," 2018 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE), Penang, 2018, pp. 263-268

[14] P. Singh, P. Pawar, P. Takwale and D. Karia, "Agriculture monitoring system AMS," 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, 2017, pp. 742-745. doi: 10.1109/ICECDS.2017.8389535

[15] L. Liu and W. Jiang, "Design of vegetable greenhouse monitoring system based on ZigBee and GPRS," 2018 4th International Conference on Control, Automation and Robotics (ICCAR), Auckland, 2018, pp. 336-339. doi: 10.1109/ICCAR.2018.8384696

[16] K. E. Adetunji and M. K. Joseph, "Development of a Cloud-Based Monitoring System Using 4Duino: Applications in Agriculture," 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD), Durban, South Africa, 2018, pp. 4849-4854. doi: 10.1109/ICABCD.2018.8465418

[17] R. N. Rao and B. Sridhar, "IoT based smart crop-field monitoring and automation irrigation system," 2018 2nd International Conference on Inventive Systems and Control (ICISC), Coimbatore, 2018, pp. 478-483. doi: 10.1109/ICISC.2018.8399118

[18] D. S. Rahul, S. K. Sudarshan, K. Meghana, K. N. Nandan, R. Kirthana and P. Sure, "IoT based solar powered Agribot for irrigation and farm monitoring: Agribot for irrigation and farm monitoring," 2018 2nd International Conference on Inventive Systems and Control (ICISC), Coimbatore, 2018, pp. 826-831. doi: 10.1109/ICISC.2018.8398915

[19] T. Hsu, D. Chang and H. Lee, "The study of application and evaluation with NoSQL databases in cloud computing," in Trustworthy Systems and their Applications (TSA), 2014 International Conference on, pp. 57-62, 2014

A Framework to Determine the Digital Skills Preparedness of Graduates for Industry

Surika Civilcharran
Information Systems & Technology
University of KwaZulu-Natal
Pietermaritzburg, South Africa
civilcharran@ukzn.ac.za

Manoj S. Maharaj
Information Systems & Technology
University of KwaZulu-Natal
Pietermaritzburg, South Africa
maharajms@ukzn.ac.zaa

***Abstract*—Rapid technological advancements are disrupting traditional job markets necessitating job seekers to develop new skillsets suitable for the digital economy. This phenomenon has a major impact on the economy, particularly in developing countries. Consequently, it is crucial for institutional offerings to be aligned to industry requirements in every discipline of higher education, in order to ultimately sustain and improve the economy. This paper presents a framework designed to determine the alignment of the digital skills that students acquire from higher education to the digital skills requirements of industry. This alignment will aid higher education institutions in improving the digital skills preparedness of their graduates, and ultimately sustaining the digital economy. Given that the digital economy requires its employees to possess a specific level of digital skills, which may vary in each sector, the proposed framework is therefore not discipline specific. Consequently, this framework may employed to establish an alignment between any discipline in higher education and the respective industry that its graduates feed into. The authors have systematically reviewed related articles to determine the factors influencing the digital skills preparedness of graduates for industry. Relevant studies were analyzed, thereby resulting in the development of a digital skills preparedness model.**

Keywords—Digital Skills, Preparedness Framework, Teaching & Learning Curricula, Institutional Challenges & Transformation

I. INTRODUCTION

Effective use of digital technology requires specific skills and talents. It is advocated that those who possess the skills to comprehend and interact with digital information to manipulate, arrange, and present it according to their understanding, possesses digital intelligence [1]. The concept of 'digital skill' has been defined in several disciplines, and as a result the terms, 'skills', 'literacy', 'competence', 'knowledge' and 'fluency' all refer to digital abilities [2]. Gekara, et al. [3] defines digital skills as "a combination of a digital mind-set (hardware, software, information, systems, security and innovation), knowledge (theoretical comprehension and understanding), competence (cognitive and practical knowhow) and attitude (value and beliefs)".

Ideally, higher education institutions prepare students with the necessary digital skills required by industry. Thereafter, these graduates find employment in the relevant sectors. Consequently, the unemployment rate drops, the country's digital economy is globally competitive and the country prospers, holistically. However, this can only be achieved if the ideologies and/or policies of government, higher education and industry are in alignment. The following authors concur that the actual circumstances are far from ideal, worldwide. Students work hard to gain a tertiary education, by taking loans and making sacrifices, in hope of attaining a job. Nevertheless, the shocking reality is that the unemployment rate of graduates in South Africa, for example, is at its highest of 453,330 (7.3%), since September 2003 [4]. After months of being unemployed, some graduates have resorted to begging for jobs at the street corners [5]. South Africa and other developing countries, face this major problems, as the growing unemployment rate negatively affects the economy, amongst other things.

In preparation for the working world, academic institutions should strive to improve their graduates' digital skills preparedness in all disciplines. Therefore, the aim of this study is to present a theoretical framework that aids in the alignment between the digital skills requirements of industry and the digital skills preparedness of graduates entering the workplace. This framework can be applied to any discipline in higher education and correlated to the respective sector within industry. For example, the digital skills requirements of organisations in the "business, real estate and finance sector" first needs to be understood. Thereafter, the digital skills of tertiary students entering this sector need to be evaluated, in order to determine the digital skills preparedness of those students. Furthermore, it is also important to explore the teaching and learning curricula of tertiary institutions regarding digital skills; the tertiary institutions policies and procedures on the digital skills of their students; as well as the challenges faced by tertiary institutions that deter institutions from efficiently and/or effectively producing digitally skilled graduates for industry.

The 'ripple effect' of employing this framework is that tertiary institutions will better understand the digital skills requirements of industry and align their curricula to meet those requirements. Consequently, graduates entering the workplace will possess the digital skillsets that is required for their sector. Additionally, these graduates will be better suited to current and future industry requirements, ultimately sustaining and improving the economy [6, 7] .

The next section discusses relevant literature that aided the development of the digital skills preparedness framework. The methodology section explains the procedure used to develop the proposed framework, and the results and discussions section provides a detailed explanation of the variables employed in the framework.

II. RELATED WORK

Information and Communications Technology (ICT) has become pervasive in contemporary societies as means of advancing educational systems, sustaining economic development via the invention of different products and services, affording access to information and proficiency to support developments in education, agriculture and health, as well as connecting communities, educators and learners [8]. In order to utilize ICT, one would not only need access to the

Internet and the technology, but also requires the digital skills necessary to use the technology effectively [8]. However, the following research suggests that there is a gap between the digital skills preparedness of graduates entering the workplace and the digital skills requirements of the various industry sectors [6, 7, 9-11].

A. The Digital Skills Crisis

The education system continuously fails to produce a cohort of young individuals who are trained to apply technology in day-to-day activities and who are prepared for the workplace once they leave the education system [10]. As a result, there is an urgent unfulfilled need for digital skills, which is predicted to get worse [10]. Therefore, there is a need for substantial and continuous investment in training and education at all levels to reduce this digital skills gap. This means that higher education institutions need to play a greater role in improving the digital skills of their students. In order to reduce the digital skills gap, it is imperative that there is an alignment between the digital skills requirements of industry and the digital skills preparedness of graduates entering the job market.

Craffert, et al. [9] highlighted that South Africa is lagging behind regarding its digital skills readiness, in contrast to the revolutionizing paradigm of a progressively digitally infused society and an increasingly digitalized economy. Due to the current digital skills shortage in industry, Craffert, et al. [9] strongly recommends that institutions of higher education prepare students with digital skills to enable them to work proficiently and effectively, thereby ensuring that future employees are digitally skilled. Furthermore, training students in multi-literacies provides students with media, information and digital skills, which helps them to associate their academic experiences to those in real-life contexts [6, 11]. Hence, universities need to restructure their degree programs, in order to incorporate the core digital skills that will prepare students for the workplace.

Craffert, et al. [9] found that the likely role of industry to guide educational institutions is being under-utilized. Subsequently, it is strongly advocated that partnerships be made between educational institutions and industry in order to guarantee that the necessary skills are supplied to sustain a digital economy [9]. At the G20 Summit, Chetty, et al. [7] proposed that the G20 requires a general digital skills support strategy, which identifies the fundamental digital skills needed by employers in developing economies, like South Africa. They explain that digital training is a core component in people harnessing enhanced usage of investments in infrastructure and technology.

B. The Impact of Digital Skills on the Economy

ICT and digital literacy is deemed an essential competence for complete involvement in an information society and a knowledge economy [11]. A South African director of Google, emphasizes that the Internet is the focal point of economic growth, and more digital skills programs are needed, which should be aimed at helping the people of Africa contribute to the digital economy [12]. Sector Education and Training Authorities (SETA) have identified skills development as a crucial requirement for South Africa's economic growth [13]. In the present highly competitive knowledge economy, there exists exponential increases of information easily available via novel technologies that are continuously evolving at an astonishing rate [6].

To be 'literate' and prolific participants in the present knowledge economy, students should graduate with the 'functional and critical thinking skills' in Information, Media and Technology that will help them to obtain a satisfying career, as well as to contribute as productive citizens to the knowledge economy [6]. He emphasizes that graduating students who do not possess these skills will be significantly disadvantaged, in a time where there is global competition for these skills.

Most developed countries have already established, or are in the process of establishing a digital economic strategy [14]. World economists, all agree that a digitally skilled workforce is necessary in order to benefit from the opportunities presented by the digital economy [14]. The European Union has successfully standardized, assessed and certified digital skills with their 'International Computer Driving License' (ICDL) course [14]. Simply being in possession of technological infrastructure, such as hardware, software and telecommunication, is insufficient to offer a competitive advantage and to prosper in the digital economy [14]. However, it is crucial to possess a workforce that can control these vital assets effectively and efficiently.

C. Digital Skills in Teaching and Learning

Over the centuries, foundational paradigms have provided frameworks that directed teaching and learning, specifically related to the Industrial Age where specialization, memorization and compartmentalization augmented productivity and profits [15]. However, in the last two decades, the pervasive infiltration of technology into daily life has transformed the way people live, in addition to the approaches used by the 21st century occupations and industries, to function [15]. In order to remain competitive, each industrial firm needed to embrace technology or fall out of the profitability pursuit [15]. This phenomenon compelled firms to introduce novel tools into their manufacturing processes and require its new employees to possess new skills [15]. This is how the 21st century fundamental skills requirements emerged [15].

Vītoliņa [16] explored the methods to enable students to employ the digital skills acquired in practice and to assess the impact of students' prior knowledge level. Her findings indicate that the skills developed are used pragmatically if the knowledge level prior to learning a new topic is higher than the average level [16]. She suggests that prior training is required as individuals reach an acceptable knowledge level. Additionally, the effect of technology enhancement positively influences higher-order thinking skills, as measured by verbal intelligence [17]. Consequently, it is imperative that each discipline, specifically in higher education, align the digital skills preparedness of its students to the respective industry sector.

Even the healthcare industry is rapidly evolving in-line with the demand for improved flexibility in the educational approach in our rapidly developing society [18]. Healthcare is undoubtedly a fast-growing industry, and graduates who choose to move in healthcare, ought to ensure they possess sound digital literacies, especially as they relate to professional communication [18]. Therefore, the continuous advancement of technologies compels student nurses to

acquaint themselves with digital skills and technologies to guarantee their use in a healthcare context is effective, safe and user-friendly [18]. It is thus vital that every discipline in in higher education align its curricular to its respective industry, as developing countries are facing the consequences of the digital skills gap.

D. Plans and Proposals for Digital Skills

Some sectors have a plan prepared to bridge the digital skills gap. For example, the skills plan for the banking sector has indicated the following implications for digital skills [10]. Firstly, there is a greater demand for advanced digital skills as a general driver of change. Since many business units are now integrated, staff need to be knowledgeable of technology across various business units and disciplines, it is therefore crucial to invest more on digital skill advancement initiatives. Additionally, new employees of the banking sector must be familiar with technology and the change in the way the banking sector now does business. Owing to technology advancements, scarce skills are no longer directly related to banking, as they now comprise process engineers, computer programmers and typically individuals in the ICT industry. Clearly, the banking sector mostly employs digital skills and these implications show the significant impact of those skills on the organisations within the sector [10].

Even as a developed country, the United Kingdom is determined to increase the digital skills of the workforce, as well as its general population [19]. For example, Higher Education includes national initiatives, such as the 'Developing Digital Literacies' programme, the 'Digital Literacies in the Disciplines' programme and the 'Changing the Learning Landscape programme' to name a few [20]. The key motivation for cultivating the digital skills of students have been acknowledged as a necessity by employers and professional bodies, as well as in Government strategies and policies [20, 21]. Chetty, et al. [7] strongly suggest that developing countries, like South Africa, require digital skills support strategies, which identifies the fundamental digital skills needed by employers in developing economies. Developing countries need to learn from the developed countries on how to bridge the digital skills gap, effectively.

III. METHODOLOGY

Numerous studies were systematically reviewed to determine the factors influencing the digital skills preparedness of graduates for industry. In total, 692 articles were assessed from which 23 articles closely met the inclusion criteria. Studies relating to digital skills of students in higher education were analyzed, thereby resulting in the development of a digital skills preparedness model (see Fig. 1).

Table 1 is the concept matrix of recent studies that influenced the development of the digital skills preparedness model. Some of the articles contributed to more than one variable. The variables identified in the concept matrix are considered core components with regard to the alignment of the digital skills required by industry to the digital skills preparedness of graduates.

TABLE 1. CONCEPT MATRIX

Variables	**Article**
Digital skills acquired through tertiary education	Chetty, et al. [22] Becker, et al. [23] Blackwood [24]
Teaching and Learning Curricula of the Institution	Kivunja [15] Vītoliņa [16] Becker, et al. [23] Funes, et al. [25]
Policies and Procedures for Tertiary Institutions	Kivunja [15] Tiatri and Jap [17] Evangelinous, et al. [20] Chetty, et al. [22] Becker, et al. [23] Davies, et al. [26] Johansson [27]
Challenges Faced by the Institution	Mosupi [12] South African Government [13] SAQA [28] Oyedemi [29] Becker, et al. [23]
Required Transformation of the Institution	Becker, et al. [23] Oliver [30] Davies, et al. [26] Ainley, et al. [8] Moloney and Farley [18]
Digital skills preparedness of students for industry	Becker, et al. [23] Alexander, et al. [31] Davies, et al. [26] Mills, et al. [32] Leckie, et al. [33] Calonge and Shah [34] Jensen, et al. [35] Brown [36]

IV. RESULTS AND DISCUSSION

A. The Proposed Digital Skills Preparedness Framework

In order to prepare students for industry demands, they need to be given access to curriculum that meets the demands of the 21st century by emphasizing the development of digital skills and competencies [37]. The institutional policies and procedures also need to be studied to ensure that they are aligned to the demands of industry. Institutional challenges ought to have a negative impact on the digital skills preparedness of students; therefore, it is essential to understand these challenges, so that appropriate measures can be taken. The institutional transformation of a university is important to improve its digital skills offerings to better suit industry requirements, and ought to have a positive impact on the digital skills preparedness of students.

Chetty, et al. [22] advocates that is imperative to recognise the correlation between the 'demand' for digital skills requirements by industry and the 'supply' of digital skills by higher education institution's training programmes. They concur that tertiary institutions must be responsive to adapt to the digital skills requirements of industry. Even though this alignment between industry and higher education is complex, it is essential for developing countries to make every effort to adapt [22]. Their study focuses more on educating learners in the school system to be adequately prepared for the workplace. However, thousands of graduates spend millions of Rand each year in developing countries, like South Africa, but are still unemployed after years of sacrifice and hard work. Consequently, more emphasis needs to be placed on finding resolutions to break this cycle soon, as people are beginning to lose faith in our education system, especially at tertiary level [22].

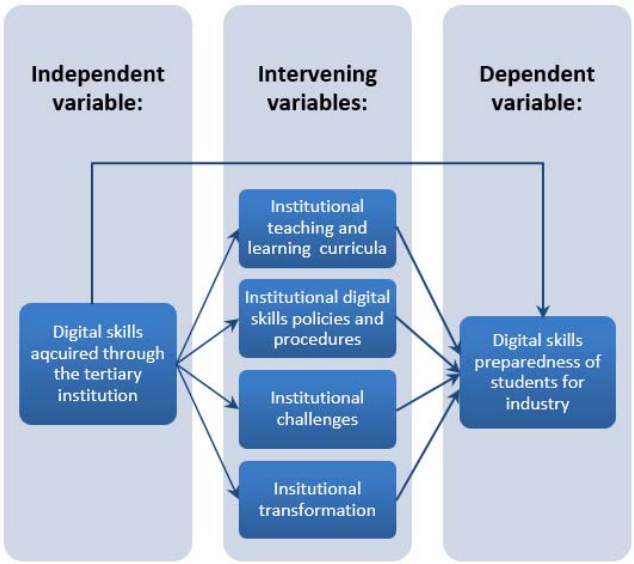


Fig. 1. Digital Skills Preparedness Model

Since each sector may require a different set of digital skills, Fig. 1 can be applied to any industry sector of any country to determine the digital skills preparedness of students entering a particular sector.

1) Independent Variable

The digital skills acquired through tertiary institutions is the independent variable of Fig. 1, as it will be manipulated to determine its impact on the dependent variable, which is the digital skills preparedness of students for industry. The dynamic digital skills requirements obtained from industry will enable higher education institutions to align their teaching and learning curricula to meet industry requirements [22]. Once these requirements are incorporated in higher education training programs, the graduates will most likely obtain formal employment in his/her field of study. This will produce a ripple effect, whereby employment opportunities would be far greater, thereby creating greater incentives for students to register for tertiary education and maybe even higher degrees [22].

2) Intervening Variables

The digital skills preparedness of students for industry is affected by a few intervening variables shown in Fig. 1, which will influence the relationship between the digital skills acquired through the tertiary institution and the digital skills preparedness of students for industry. The significance of each intervening variable in relation to the digital skills preparedness of graduates for industry, is discussed below.

a) Teaching and Learning Curricula of the Institution

The teaching and learning curricula of an institution (i.e. the first intervening variable in Fig. 1) plays a pivotal role in the digital skills preparedness of students for industry. The growing demand for the labour force to be digitally competent coerces higher education institutions to evaluate their curricula and guarantee that digital skills are embedded as a graduate trait [20, 21]. In order to produce students equipped with digital skills to help them become proficient in the 21st century globalized trades, professions and occupations, it is crucial that digital technologies are embedded in pedagogy and curriculum [15, 20]. After studying the learning paradigms that have directed pedagogy over time, Kivunja [15] asserts that a shift in curriculum and pedagogy is needed, towards a paradigm that highlights problem solving and critical thinking. Typically, digital skills are developed by teaching and learning, therefore, it is imperative to offer such hands-on training on digital skills, particularly if individuals gain from acquiring new skills [16]. For learning of digital skills, a variety of suitable learning resources must be provided, which the learner can choose to read or watch several times, subject to the student's needs and experience [16]. Brown [36] maintains that digital skills are necessary for survival, production and creativity, embracing digital and open resources, learning through virtual and physical environments, and research in real-time.

b) Policies and Procedures for Tertiary Institutions

Policies and procedures adopted by an institution (i.e. the second intervening variable in Fig. 1) should aim to align the teaching and learning curricula to the digital skills requirements of industry. Establishing institutional quality-control policies and procedures will guarantee that students graduate with the essential digital skills required in their particular field [20]. Tiatri and Jap [17] maintains that the effect of technology enhancement positively influences higher-order thinking skills, as measured by verbal intelligence. The development of digital skills relative to their discipline offers the potential of inspiring graduates to become digital leaders and innovators within their disciplines [20]. Furthermore, measuring digital literacy and regularly guaranteeing that policies and procedures are adequately agile to respond to dynamic nature of digital skills' will lead to efficiency across a country [22]. In order to provide students with an effective education in the 21st century, a pedagogical paradigm shift is necessary so that students are prepared with the skills required in the 21st century workplace, and must be also proficient in problem solving, thinking for themselves, working in teams and leading others to triumph in the Digital/Knowledge Economy [15].

c) Challenges Faced by the Institution

Some challenges faced by an institution (i.e. the third intervening variable in Fig. 1) ought to impede the digital skills preparedness of students, therefore it is considered an intervening variable. It is imperative that these institutional challenges are identified, so that the South African Department of Higher Education can present viable solutions. The South African Qualifications Authority (SAQA) have not catered for the development of digital skills in their South African National Qualifications Framework [28]. South Africa acquiring the technology and infrastructure from developed countries does not imply that the users will possess the necessary digital skills to make effective use of the investments in the technology and infrastructure by the government and industry [12]. Furthermore, existing cross-national studies of digital skills have largely been undertaken in developed countries with greater levels of ICT resources and ICT development [13].

Consequently, as a developing country, South Africa needs to determine if university students possess the core digital skills required by industry [12]. University students are predominantly relevant to the study of digital skills, since most South African universities provide students with access to computer laboratories, and studies show that education is a major indicator of digital skills [29]. It is thus necessary to understand the challenges faced by tertiary

institutions, so that correctional measures can be developed/taken to ensure that the teaching and learning curriculum, as well as policies and procedures cater for the digital skills upliftment of tertiary students to ultimately produce a digitally skills workforce.

d) Required Transformation of the Institution

The required institutional transformation (i.e. the fourth intervening variable in Fig. 1) is essential in order to better align the institution's teaching and learning to meet the digital skills requirements of industry. In order to understand the degree and nature of institutional transformation required, it is important to first establish the following, respectively:

- the teaching and learning curricula on digital skills;
- the policies and procedures of tertiary institutions on digital skills;
- the challenges faced by tertiary institutions that deter them from efficiently and/or effectively producing digitally skilled graduates.

The transformation of the institution is necessary if its digital skills alignment with industry is poor or non-existent. The institutional transformation will deal with strategies to: (1) improve the teaching and learning curricula; (2) update the policies and procedures; and eliminate the challenges face by the institution.

3) Dependent Variable

Essentially, the degree of digital skills acquired through tertiary institutions will explain the variance in the digital skills preparedness of students for industry. The European Commission advocated that 90% of jobs may soon require basic digital skills in accountancy, medicine, architecture, engineering, art, nursing and many more careers [19, 38]. The UK, as a first world country, has identified the impact of the digital revolution on the country and has created reports/policies to help its citizens improve their digital skills. This in turn will benefit UK's economy, holistically. Research suggests that skill-biased technological change is rapidly transferred from the developed countries to the developing countries [39]. Therefore, South Africa, as a developing country, needs to follow suit or else our economy will continue to suffer.

The difference between the dependent and independent variable in Fig. 1 is that the independent variable is the digital skills that students acquire through tertiary institutions. However, the dependent variable is the digital skills knowledge that is required at industry level. From a primitive perspective the digital skills acquired by students at tertiary institutions in South Africa are designed to have a strong alignment to industry expectations (as illustrated in the direct relationship between the independent and dependent variables in Fig 1). However, a more informed examination of this relationship reveals a significant contribution by a set of intervening variables that also has an influence on the relationship between the independent and dependent variables (also illustrated in Fig.1).

V. CONCLUSION

Technology advancements necessitates graduates entering the workplace to possess the necessary digital skills. Thus, it is vital that digital skills development of tertiary students, as well as infrastructure development be given the same priority, if these graduates entering the industry sectors are to compete in the digital economy. Furthermore, international educators, government, organizations and institutions all see the need to develop and implement plans to improve the digital skills of students and the workforce, with the aim of sustaining the economy. In response to the issues discussed, the proposed digital skills framework will assist researchers in bridging the gap between industry requirements and academia curricula concerning digital skills. The framework may be applied to each discipline in order to align its teaching and learning curricula, as well as its institutional policies and procedures to the digital skills requirements of its respective industry.

VI. RECOMMENDATIONS

Further research may be conducted to test the proposed framework using different methodologies. Additionally, this framework may be used to investigate the digitals skills requirements in various other disciplines in order to investigate the digital skills preparedness of students entering other industry sectors. Once tested, the digital skills preparedness model may be used as a theoretical framework as the basis of future studies.

ACKNOWLEDGMENT

The authors declare that they have no financial or personal relationships which may have inappropriately influenced them in writing this article. This article, resulted from the doctoral research of S.C. and M.S.M. provided editorial oversight and final approval.

REFERENCES

[1] N. B. Adams, "Digital intelligence fostered by technology," Journal of Technology Studies, vol. 30, pp. 93-97, 2004.

[2] E. M. Meyers, I. Erickson, and R. V. Small, "Digital literacy and informal learning environments: an introduction," Learning, Media and Technology, vol. 38, pp. 355-367, 2013.

[3] V. Gekara, A. Molla, D. Snell, S. Karanasios, and A. Thomas, "Developing appropriate workforce skills for Australia's emerging digital economy," RMIT University, NCVER, Adelaide, 2017.

[4] Stats SA. (2017). Key findings: P0441 - Gross Domestic Product (GDP), 2nd Quarter 2017. Available: http://www.statssa.gov.za/?page_id=1856&PPN=P0441&SCH=6983

[5] N. Mngoma. (June 13, 2016). Graduates are 'Begging' for Jobs. Available: https://www.iol.co.za/news/opinion/graduates-are-begging-for-jobs-2034214

[6] C. Kivunja, "Unpacking the information, media, and technology skills domain of the new learning paradigm," International Journal of Higher Education, vol. 4, pp. 166-181, 2015.

[7] K. Chetty, U. Aneja, V. Mishra, N. Gcora, and J. Josie, "Bridging the Digital Divide: Skills for the New Age," G20 Summit Germany, 2017.

[8] J. Ainley, W. Schulz, and J. Fraillon, "A global measure of digital and ICT literacy skills," Australian Council for Educational Research ED/GEMR/MRT/2016/P1/4, 2016.

[9] L. Craffert, M. Ungerer, K. Visser, and J. Morrison, "Strategies, practices and skills for competitiveness in the digital economy: a perspective on large companies in South Africa," in Proceedings of the e-Skills for Knowledge Production and Innovation Conference, Cape Town, South Africa, 2014.

[10] A. Schofield, "JCSE ICT skills survey," Joburg Centre for Software Engineering, 2016.

[11] A. Klimczuk, J. Jasiewicz, M. Filiciak, A. Mierzecka, M. Kisilowska, A. Tarkowski, et al., "Framework directory of digital skills," Polska, Warszawa, 2015.

[12] A. Mosupi, "Google to train 1-million young Africans in digital skills," in Times Live, ed: Times Media Group, 2016.

[13] South African Government. (2016, 26 January 2017). About SA education. Available: http://www.gov.za/about-sa/education

[14] C. Chinien and F. Boutin. (2011, November 13, 2017). Defining essential digital skills in the canadian workplace: final report Available: http://en.copian.ca/library/research/digi_es_can_workplace/digi_es_can_workplace.pdf

[15] C. Kivunja, "Do you want your students to be job-ready with 21st century skills? Change pedagogies: a pedagogical paradigm shift from Vygotskyian social constructivism to critical thinking, problem solving and siemens' digital connectivism," International Journal of Higher Education, vol. 3, pp. 81-91, 2014.

[16] I. Vītoliņa, "E-inclusion process and societal digital skill development," Discourse & Communication for Sustainable Education, vol. 6, pp. 86-94, 2015.

[17] S. Tiatri and T. Jap, "Preliminary investigation on the effectiveness of a thinking skill training in Indonesia: "thinking skills training with digital technology"," Journal of Psychological & Educational Research, vol. 23, pp. 41-53, 2015.

[18] C. Moloney and H. Farley, "Digital skills in healthcare practice," in Building professional nursing communication, ed. Port Melbourne, Australia: Cambridge University Press, 2015, pp. 155-181.

[19] D. Lill. (2015) Digital skills essential for future-proofing career. © Talkbusiness Magazine. Available: http://www.talk-business.co.uk/2015/11/05/digital-skills-essential-for-future-proofing-career/

[20] G. Evangelinous, D. Holley, and M. Kerrigan, "Implementing a model and processes for mapping digital literacy in the curriculum (online badges)," in Proceedings of the European Distance and E-Learning Network 2016 Annual Conference, Budapest, 2016.

[21] Quality Assurance Agency, "Higher education review: themes for 2015-16," Gloucester, UK, 2014.

[22] K. Chetty, Q. Liu, L. Wenwei, J. Josie, N. Gcora, and B. Shenglin, "Bridging the digital divide: measuring digital literacy," G20 Germany, 2017.

[23] S. A. Becker, M. Cummins, A. Davis, A. Freeman, C. G. Hall, and V. Ananthanarayanan, "NMC horizon report: 2017 higher education edition," The New Media Consortium, United States, 2017.

[24] N. Blackwood, "Digital skills crisis: second report of Session 2016–17," House of Commons HC 270, 2016.

[25] J. C. M. Funes, F. Aguirre, F. Deeg, and J. Hoefnagels, "Skills for tomorrow: How to address the digital skills gap," The Future of Work, vol. Policy Brief Series, pp. 1-5, 2018.

[26] S. Davies, J. Mullan, and P. Feldman, Rebooting learning for the digital age: What next for technology-enhanced higher education?: Higher Education Policy Institute Oxford, 2017.

[27] J. Johansson, "Rethinking Education–A Critical Analysis of the Increased Emphasis on Digital Skills in EU (Higher) Education Policy," in Swedish Political Science Association (SWEPSA), Karlstad, Sverige, 2017.

[28] SAQA, "General and further education and training qualifications sub-framework and higher education qualifications sub-framework of the national qualifications framework," vol. 578, ed. Pretoria: South African Qualifications Authority, 2013.

[29] T. Oyedemi, "Beyond access: youth and digital skills," Communicatio, vol. 40, pp. 137-154, 2014.

[30] B. Oliver, "Redefining graduate employability and work-integrated learning: Proposals for effective higher education in disrupted economies," Journal of Teaching and Learning for Graduate Employability, vol. 6, p. 56, 2015.

[31] B. Alexander, S. A. Becker, M. Cummins, and C. H. Giesinger, "Digital literacy in higher education, Part II: An NMC Horizon project strategic brief," The New Media Consortium, United States, 2017.

[32] A. Mills, A. K. Sanders, and S. Hussain, "Digital Skills Implementation in Journalism and Mass Communication Curriculums: A Census of Programs in the Middle East," in Qatar Foundation Annual Research Conference Proceedings, 2018, p. SSAHPD1084.

[33] N. Leckie, J. Rodier, and D. Gyarmati, "Workplace digital essential skills in rural small businesses: final research report and highlights," Social Research and Demonstration Corporation (SRDC), Canada, 2016.

[34] D. S. Calonge and M. A. Shah, "MOOCs, graduate skills gaps, and employability: A qualitative systematic review of the literature," The International Review of Research in Open and Distributed Learning, vol. 17, pp. 67-90, 2016.

[35] H. V. Jensen, S. Tofalleti, C. Thornley, S. Murnane, Grand Coalition for Digital Jobs, and Digital Europe, "Digital Skills for SMEs: get inspired now!," Digital Europe, Brussels, Project Report, 2016.

[36] B. Brown, "Twenty first century skills: a Bermuda college perspective," Voices in Education: Journal of Bermuda College, vol. 1, pp. 58-64, 2015.

[37] Y. Rosen, Handbook of research on technology tools for real-world skill development: IGI Global, 2015.

[38] A. TurculeŢ and E. MihuŢ, "A comparative approach: digital skills in formal teaching and learning," presented at the The 11th International Scientific Conference eLearning and Software for Education, Bucharest, Romania, 2015.

[39] M. Vivarelli, "Innovation, employment and skills in advanced and developing countries: a survey of economic literature," Journal of Economic Issues, vol. 48, pp. 123-154, 2014.

The Perception of Students about the use of Social Media as an Alternate Learning Platform

Sindisiwe Mahlambi
Information Systems & Technology
University of KwaZulu-Natal
Pietermaritzburg, South Africa
212558630@stu.ukzn.ac.za

Surika Civilcharran
Information Systems & Technology
University of KwaZulu-Natal
Pietermaritzburg, South Africa
civilcharran@ukzn.ac.za

Nurudeen A. Ajayi
Information Systems & Technology
University of KwaZulu-Natal
Pietermaritzburg, South Africa
ajayi@ukzn.ac.za

***Abstract*—The advent of Web 2.0 brought about a new medium of communication, such as social media. Today's generation of students have been greatly influenced by the Web and hence, have become active users of the different available social media platforms. The influence of social media platforms on students, is causing a shift in learning, and as a result, educational institutions are increasingly creating alternative learning platforms, using social media to facilitate student-centred learning. This paper presents the findings from a study that was conducted to understand the perception of students regarding the use of social media as an alternative learning platform. The quantitative research method was adopted, and 331 questionnaires were administered at the Pietermaritzburg campus of University of KwaZulu-Natal (UKZN). The findings reveal that students are comfortable with the use of social media as a learning platform, because it enhances their ability to be interactive and it improves their chances of accessing educational information, even when not within the academic environment. The findings, however, also revealed that some features of social media could cause distraction to students, but when adequate measures are introduced, social media could become an additional means of interaction for students, especially for academic purposes.**

Keywords—Social Media, Information Technology, Teaching and Learning Platforms, Higher Education

I. INTRODUCTION

The evolution of the digital landscape has brought about a change in the culture of academia. With easy access to smart devices, students have become digitally influenced. Moreover, the Internet is considered a highly valuable tool, with YouTube, Facebook and Twitter being among the highly trafficked social media platforms [1]. Social media platforms have become popular in universities. For example, a study at the University of Cape Town found that students preferred Facebook to their course management system [2]. Some American and Canadian universities are also utilising social media platforms to share lectures, tutorials, and even research data [3].

The pedagogical value of social medial platforms is in their ability to act as a catalyst in enhancing teaching and learning in Higher Education [4-7]. The popularity of social media platforms is undeniable. For example, Facebook has more than 900,000,000 estimated monthly visitors, and Twitter about 310,000,000 estimated monthly visitors, [8]. Even though face-to-face interaction offers many advantages for teaching and learning, it is limited by space and time, [9]. Social medial platforms, on the other hand, seems to cater for the limitation of space and time that exist in face-to-face interaction, by enabling teaching and learning interaction to occur at any time of the day and at the comfort of both the educator and the student. Social medial platforms also seems to capture students' attention beyond the classroom to revise what they have learnt or will learn in subsequent lectures, [9]. Given the extraordinary statistics of social medial platforms and their increasing adoption in academia, it has become important to understand the perception of students about social medial platforms as a tool in enhancing their studies. In light of this, this study was conducted to understand the perception of students about the use of social media as an alternative learning platform. This study also explored the challenges that could arise from adopting social media as a learning platform, as well as measures to manage those challenges.

II. LITERATURE REVIEW

In recent times, technology has played a major role in academia by introducing new methods of teaching and learning, [6]. Teenagers today, are commonly known as "digital natives. At tertiary level, students are now largely dependent on Web tools, such as Learning Management Systems (LMS) and search engines to assist them with learning.

Learning methods have changed over time and much research has been conducted in this area, but researchers and practitioners have still not agreed on common definitions of terms, such as online learning and e-learning. As a result, these terms are often used interchangeably [10]. While some of the characteristics overlap between the two concepts, there are a few differences, which may imply that the concepts should not be used interchangeably. For this reason, it is necessary to first distinguish between online learning and e-learning.

A. Online Learning vs. E-Learning

Online learning is considered the most difficult to define, but most researchers broadly describe it as the use of technology to access learning experiences. Moore et al. [10] describes both e-learning and online learning as a sub-level of distance learning that uses Web tools and the Internet. Furthermore, these tools allow learners to access learning on campus or off-campus. To allow for effective learning, for both e-learning and online learning, the learning environment must have interactive tools that allows for efficient correspondence between the learner and the educator.

However, researchers are having conflicting views on the definition of e-learning and online learning. The differences between e-learning and online learning as explained by Moore et al. [10] is that e-learning combines multimedia or electronic devices, and is blended with face-to-face learning. In contrast, online learning may be partially online/hybrid or completely online with no face-to-face learning present. Based on the aforementioned definitions and a study by Qureshi et al. [11], social media can be categorised as an e-learning tool. E-Learning has been further categorised into synchronous or asynchronous [11, 12].

1) Asynchronous E-Learning

Asynchronous e-learning offers connectivity between educators and learners, even when learners or educators are not present [11]. It also allows educators to share information and resources with learners in absentia. Asynchronous e-learning is mainly facilitated by media tools, such as e-mail and discussion forums and is considered a key factor in flexible e-learning [12].

2) Synchronous E-Learning

Synchronous e-learning facilitates collaboration and participation at the same time [11]. Synchronous e-learning, mainly facilitated by media tools, such as chat and videoconferencing, is considered to have the ability to support e-learners [12]. Therefore, social media can be categorised as synchronous e-learning [11].

B. Social Media for Learning Purposes

Web 2.0 is the second generation of the World Wide Web, which enables collaboration and information sharing by users. It was initially developed for non-educational purposes, but has since, brought about the concept of social media [13]. Web 2.0 technologies are making a major impact on communicative behaviours. They are also causing a remarkable growth in the popularity of Social Networking Sites (SNS) such as Facebook and YouTube, focusing on collaboration and social activities. Tertiary institutions are realising the advantages of using Web 2.0 technologies for learning, especially since most students are increasingly becoming proficient with the use of these technologies [7]. Most tertiary institutions are, therefore recognising that these technologies can be utilised to develop periodically sound environments that would be useful to both students and lecturers [14].

Today, students while doing their homework are often observed instant messaging, watching YouTube, using Facebook, Twittering, and surfing the Web at the same time [9]. Students that belong to this generation, without instruction, develop their own cognitive skills that are needed for enquiry-based learning, network learning, discovery-based learning, collaborative learning, experimental learning, active learning, self-regulation and organisation, and problem solving [15].

Some of the ways that today's teens spend their time has changed compared to the past, but early adulthood and the adolescence central developmental tasks, which are the phases that people experience when growing up, have remained [16]. Formation of identity, power of the peer group, and development of intimate relationships are some of the key characteristics of adolescent development. Today, most young adults, especially students, experience most of these key characteristics through social networking websites like Facebook [16].

Social media platforms have very distinct application features, such as groups for interaction on common subjects, wall posts, posting status messages, keeping in contact with friends through chatting, and there is a wide opportunity to use these features for the purpose of facilitating learning. The platforms can facilitate academic discussions, help students toward their university courses, as well as in gaining information and establishing interaction with their peers [11, 17]. Using the platforms, students can freely share knowledge and information about their studies.

Bosch [2] found that student's use of social media is more extensive than their use of course administration system offered by the University of Cape Town. Furthermore, undergraduate students were found to be heavier users of social media, whereas postgraduate students used social media more for networking rather than social engagement. Social media has been welcomed on its use for academic purposes, because studies show that students indicated that they had already been spending a significant amount of time social media platforms, so having the option to check their course related content was considered very easy. Studies also indicated that students, through social media, assisted each other in locating learning materials and in answering questions. Additionally, they kept in touch during holiday periods to discuss holiday projects and share study and lecture notes through social media platforms.

As critical thinking is one of the objectives of most tertiary institutions, it can be said that social media encourages critical thinking by allowing students to interact with their peers taking the same courses [18]. It is believed that social media has the ability to drastically change the educational system by motivating students to be more active rather than just being observers in educational practices [11, 19]. According to Gray [18], students should keep an open mind about using social media for educational purposes, as this is the way in which it can have a positive influence on their studies.

Blended learning is a combination of traditional teaching and learning, which involves face-to-face interaction between educators and students, and e-learning. Using social media for the purpose of learning could provide the benefit of blended learning, which in turn, provides a positive experience to both students and educators [20]. Shih [20] found that students had a positive attitude toward the e-learning approach on social media, as compared to the traditional learning approach. Dzvapatsva et al. [9] concluded that social media platforms offers educators an alternative method for teaching and learning, and increases learning time. Mbatha [7], however, strongly recommends that additional research is required in order to further understand how Web 2.0 may effectively be adopted to enrich teaching and learning.

1) Social Media as a Support Mechanism

In 2010, an earthquake hit New Zealand, in the Canterbury region. This unfortunate event led to the closure of the University of Canterbury for two weeks, and the university relied on their website during this time to inform students of events, activities and progresses made with regards to the rebuilding and reopening of the school(s). A Facebook community was established for the university, enabling continuous interaction and information sharing between the students and university staff, as well as supporters [17]. This community was well accepted, and passionately used by students, staff members and well-wisher of the university. This was one of the events that led people to believing that social media platforms, such as Facebook can provide a vehicle for the sharing and accessing of critical information [17]. According to Dabner [17], a staff member at the university, he could not gain access to the physical facilities of the university during its closure after the earthquake, so he utilised Facebook and the campus website daily to get up-to-date information during the crisis. Disasters and crises are inevitable and the introduction of

technology innovations and an increase in the use of social media has provided a new method for people to communicate during such times. This suggests that social media can extensively aid students during disaster periods, and the inaccessibility of an LMS.

2) Pros and Cons of Integrating Learning into Social Media

As with every learning platform, so too does social media have it benefits and drawbacks. There have been conflicting results from various studies on using social media in an academic context. Some studies found that the use of social media by students impacted their study time resulting in poor academic results, whereas other studies found no such evidence of this [21]. According to Irwin et al. [21], the previous studies that had these conflicting results considered only the effects of time management on social media on the academic performance of students, rather than focusing on the overall potential educational benefits that Facebook could provide students.

a) Benefits of Integrating Learning into Social Media

Social media can be seen as an informal learning environment that facilitates extended interaction and learning among peers [9]. With all the available functions, social media has the potential to establish spontaneous relationships, and further support interactive learning and communication. With the ability to create groups and make comments, SNS attracts students to take part in learning gradually, making the learning process more successful and effective, and providing an opportunity for students to acquire new knowledge [4].

Wang et al. [22] studied students' perceptions of using a Facebook group as an LMS and found students to be in agreement with the usage of the Facebook group as being successful. Undergraduate students felt like they got to know their peers better and had the opportunity to make friends with people doing the same course [22]. It can thus, be argued that social media helps students to establish relationships with their peers, and this includes academic relationships. Social media also has the potential to be used as an LMS, since it allows the sharing of resources, making of announcements, online discussions and student collaboration [18, 20, 22].

b) Drawbacks of Integrating Learning into Facebook

Students may at times find it difficult to draw the line between using social media for socialising and using it for learning. For this reason, it requires group effort to maintain a balance between learning and socialisation on social media, and may be better if social media groups, created for the purpose of a university course, was administered, guided and monitored by at least a lecturer [18]. However, the problem with university staff intervening would be that students see social media beyond the control of their lecturer(s) as it is very different form a classroom or academic website setting, and it would be difficult for a staff member to administer control of their students on a particular group on the site [18].

Wang et al. [22] in their study found that even though Master's students agreed that the social media group helped them with peer communication, it did not create close relationships and merely forced interaction between them, which did not work out well. Hence, the Master students' opinions were found to be more negative compared to undergraduate students, that believed that social media is helping them create a relationship [22]. Using social media in a higher education environment is not free from potential limitations and risks. Stalking, privacy, content ownership, virtual integrity and cyber bullying are issues that must be taken into consideration.

III. METHODOLOGY

This study adopted a quantitative research approach, and was conducted at the University of KwaZulu-Natal (UKZN), Pietermaritzburg (PMB) campus, South Africa. The target population were undergraduate students within the School of Management, Information Technology & Governance. The target population was chosen based on convenience and proximity to the researchers. A non-probability and purposive sampling technique was used, hence, participants were selected based on the fact that they are familiar with and have a social media profile. Another important factor considered in the selection of participants was that, they must have used at least one online learning platform for academic purposes. This is important, because in the questionnaire respondents were expected to give an account of their experience with the use of social media for academic purposes. Questionnaires were distributed to 331 students, but only 113 questionnaires out of those returned were deemed fit for use in the study. In addition to a low response rate, many questionnaires were incomplete. Consequently, this would have adversely affected the overall findings if incomplete responses were included. The statistical analysis was performed on collected data, using SPSS.

IV. FINDINGS

The findings of this study, based on the analysis and interpretation of the analysed data, are presented below.

A. Ease of Use

Responses from the participants indicate that student find social media platforms easy to use for learning purposes. This is because of their continuous exposure and use of social media platforms. In the work of Pempek et al. [16], it was presented that social media has become part of many students' daily lives and hence, most students spend at least 30 minutes daily on at least one of its platforms. This also aligns with the findings of Bosch [2], where it was found that students preferred social media platforms to their course management system, owing to their familiarity with social media, and its ease of use.

Most participants in this study stated that they often forget to check the Learning Management System (LMS) used by the University for teaching and learning purposes. This, according to them, is because they became exposed to the LMS only after gaining admission into the university. However, with social media, most student indicated that they have had their accounts and hence have grown used to it even before registering as a university student, and as such, use their social media account regularly.

Participants identified instant messaging, group chats and voice and video call features, as components of social media that they have and will continue to use for academic purposes. As shown in Fig. 1, 70% of respondents indicated that, on several occasions, they have used their walls on

Facebook to post information, such as website links to academic resources, as well as changes in dates of assignment submission and assessments. They also indicated that they have used the instant messaging function of other social media platforms, like WhatsApp to pass urgent information, such as impromptu change in lecture venues and cancellation of lectures, because they believe most of their peers have their WhatsApp chat opened on their mobile devices and, hence will be able to receive these messages instantaneously. The questionnaire responses also showed 75% of respondents are members of academic groups on WhatsApp and Facebook, as they believe that they can relate to people with common topical interests that are in the different groups they might have joined.

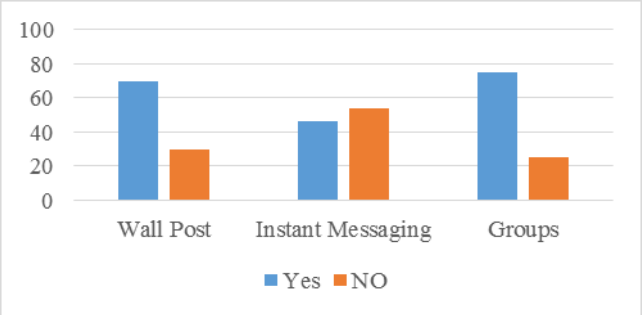


Fig. 1. Social Media Features often used for Academic purposes.

B. *Distance Learning*

Some questions in the questionnaire was used to understand how social media can be relevant and used for academic purposes. The responses from the participants show that social media can be used as a platform to support distance learning and also facilitate e-learning. Participants were asked if they considered social media platforms, such as WhatsApp and Facebook as a platform that can provide a flexible environment for learning, and still encourage students' participation. 54% of the respondents indicated that the use of social media platforms will facilitate a learning environment where they can participate in academic activities extensively, while the remainder, however, believes that it will not. There seems to be a slight margin in the difference, but what can be deduced is that more respondent believe social media can encourage student participation in academic activities and also facilitate a responsive learning environment, in which students do not have to be physically present in the classroom. This indication of respondents in this study, is in agreement with the work of Qureshi [11], in which social media has been categorised as an e-learning tool that aids distance learning and blended learning.

C. *Enhances collaboration*

Owing to their familiarity with different social media platforms, participants believe that social media features, such as chatrooms, groups and video calls could serve as effective means of establishing academic interactions and collaboration amongst themselves. This is because these features allow students to post information relevant to their course, join groups where they can share ideas and collaborate in order to find solution to assessment tasks and if necessary be able to chat instantly with their peers. Al-Rahmi and Zeki [23] concurs that social media tools has a positive influence in enhancing learner performance and promotes active collaborative learning for students in higher education.

Additionally, social media, through features, such as video chats allow people to connect and interact with each other virtually. Hence, 79% of the respondents considered social media as one of the most popular and effective means of facilitating collaboration among students. Some responses, however, showed that these features could cause distractions to students, and that students may find it difficult to manage their academic and social activities on social media platforms.

D. *Managing students academic activities on Social Media*

As stated in section C, the findings of this study revealed that some features of social media could cause distraction to students, and as such, students could find it difficult to manage and distinguish between academic activities and their social activities, especially when expected to complete or participate in academic tasks or activities on a social media platform. As found from responses in this study, measures to manage the challenges faced by using social media as a learning platform includes having a time limit on activities and having guidelines on interactions related to academic activities. Watson [24] found that the more social media distracted students, the more they lost academic success qualities, and obtained academic frustration qualities. In order to prepare undergraduates for a successful academic career, techniques for traversing digital interactions ought to be established for incoming students [24].

V. CONCLUSION

There are many factors, such as ease of use and familiarity, which are influencing the adoption of social media as a learning platform. These factors, as shown by the participants of this study enhances the use of social media platforms as an e-learning tool, which in turn enhances collaboration among students and also facilitates distance learning. Previous studies on students' use of LMS shows that student only tend to use their University's LMS if they are forced to use it, unlike with social media platforms which have become part of their everyday life.

Since most students are familiar with social media, it different platforms and its features, they seem to prefer using it as a LMS. It has therefore become essential to understand how social media shapes academic relationships and the collaboration of diverse students, especially in African environments [1]. This study shows that the relationships that students keep through social media is an important factor that could pave the way to them to regularly interact among themselves and with their lecturers. However, it should be noted that social media should not be considered as a complete replacement of the different available LMSs, but its incorporation as an alternative or complementary learning platform at universities may have a positive impact on student's course delivery satisfaction.

VI. RECOMMENDATION

Future research should be conducted by applying the Technology Acceptance Model with a larger sample population at two separate academic institutions. The results emanating from both these clusters can then be compared by statistically analysing the two data sets, using T-tests. The

T-tests will inform the researcher the mean of each cluster, the average difference among the clusters, and if the difference is statistically significant. This research approach will result in a better understanding of the concepts under investigation.

ACKNOWLEDGMENT

The authors declare that they have no financial or personal relationships, which may have inappropriately influenced them in writing this article.

REFERENCES

[1] P. Rambe and D. Ng'ambi, "Learning with and from Facebook: Uncovering power asymmetries in educational interactions," Australasian Journal of Educational Technology, vol. 30, pp. 312-325, 2014.

[2] T. E. Bosch, "Using online social networking for teaching and learning: Facebook use at the University of Cape Town," South African Journal for Communication Theory and Research, vol. 35, pp. 185–200, 2009.

[3] C. Fernàndez and E. P. Gil-Rodríguez, "Facebook as a collaborative platform in higher education: The case study of the universitat oberta de Catalunya," in Technology-Enhanced Systems and Tools for Collaborative Learning Scaffolding, ed: Springer, 2011, pp. 27-46.

[4] B. Chen and T. Bryer, "Investigating instructional strategies for using social media in formal and informal learning," The International Review of Research in Open and Distributed Learning, vol. 13, pp. 87-104, 2012.

[5] L. Deng and N. J. Tavares, "From Moodle to Facebook: Exploring students' motivation and experiences in online communities," Computers & Education, vol. 68, pp. 167-176, 2013.

[6] M. Graham, "Social media as a tool for increased student participation and engagement outside the classroom in higher education," Journal of Perspectives in Applied Academic Practice, vol. 2, pp. 16-24, 2014.

[7] B. Mbatha, "Global transition in higher education: From the traditional model of learning to a new socially mediated model," The International Review of Research in Open and Distributed Learning, vol. 15, pp. 257-274, 2014.

[8] Ebiz|MBA. (30 March 2017). Top 15 Most Popular Social Networking Sites. Available: http://www.ebizmba.com/articles/social-networking-websites

[9] G. P. Dzvapatsva, Z. Mitrovic, and A. D. Dietrich, "Use of social media platforms for improving academic performance at Further Education and Training colleges," South African Journal of Information Management, vol. 16, pp. 1-7, 2014.

[10] J. L. Moore, C. Dickson-Deane, and K. Galyen, "e-Learning, online learning, and distance learning environments: Are they the same?," The Internet and Higher Education, vol. 14, pp. 129-135, 2011.

[11] I. A. Qureshi, H. Raza, and M. Whitty, "Facebook as e-learning tool for higher education institutes," Knowledge Management & E-Learning, vol. 6, pp. 440-448, 2014.

[12] S. Hrastinski, "Asynchronous and synchronous e-learning," Educause quarterly, vol. 31, pp. 51-55, 2008.

[13] V. Benson, Cutting-edge technologies and social media use in higher education: IGI Global, 2014.

[14] R. M. English and J. A. Duncan-Howell, "Facebook© goes to college: Using social networking tools to support students undertaking teaching practicum," Journal of Online Learning and Teaching, vol. 4, pp. 596-601, 2008.

[15] P. A. Kirschner and A. C. Karpinski, "Facebook® and academic performance," Computers in human behavior, vol. 26, pp. 1237-1245, 2010.

[16] T. A. Pempek, Y. A. Yermolayeva, and S. L. Calvert, "College students' social networking experiences on Facebook," Journal of applied developmental psychology, vol. 30, pp. 227-238, 2009.

[17] N. Dabner, "'Breaking Ground'in the use of social media: A case study of a university earthquake response to inform educational design with Facebook," The Internet and Higher Education, vol. 15, pp. 69-78, 2012.

[18] K. Gray, L. Annabell, and G. Kennedy, "Medical students' use of Facebook to support learning: Insights from four case studies," Medical teacher, vol. 32, pp. 971-976, 2010.

[19] N. Selwyn, "Screw Blackboard... do it on Facebook!': an investigation of students' educational use of Facebook," Ponencia. En: Poke, 2007.

[20] R.-C. Shih, "Effect of using Facebook to assist English for business communication course instruction," Turkish Online Journal of Educational Technology-TOJET, vol. 12, pp. 52-59, 2013.

[21] C. Irwin, L. Ball, B. Desbrow, and M. Leveritt, "Students' perceptions of using Facebook as an interactive learning resource at university," Australasian Journal of Educational Technology, vol. 28, 2012.

[22] Q. Wang, H. L. Woo, C. L. Quek, Y. Yang, and M. Liu, "Using the Facebook group as a learning management system: An exploratory study," British Journal of Educational Technology, vol. 43, pp. 428-438, 2012.

[23] W. M. Al-Rahmi and A. M. Zeki, "A model of using social media for collaborative learning to enhance learners' performance on learning," Journal of King Saud University-Computer and Information Sciences, vol. 29, pp. 526-535, 2017.

[24] C. Watson, "Social media distractions effect traits differently across generations," presented at the Idaho Conference on Undercraduate Research, Boise, Idaho, July 25-26, 2018.

Electrical Energy Usage Analysis for all Residences of a South African Academic Institution

Sula B. K. Ntsaluba
Dept. of Elecrical and Electronic Engineering Technology
University of Johannesburg
Johannesburg, South Africa
ntsaluba.sbk@gmail.com

Maria Michael
Dept. of Electrical and Electronic Engineering
University of Pretoria
Pretoria, South Africa
maria.michaelm@gmail.com

Solathiso Shasha
Dept. of Electrical and Electronic Engineering
University of Pretoria
Pretoria, South Africa
solathisoshasha@gmail.com

Isaac Dlamini
Dept. of Electrical and Electronic Engineering
University of Pretoria
Pretoria, South Africa
isaacdlamini3802@gmail.com

***Abstract*—This study presents an analysis of the energy usage of the twenty-eight residences at the University of Johannesburg during the 2017 calendar year. The paper presents the energy consumption if the residences based on the total energy usage per residence as well as in terms of the number of students at each of the residences for the different seasons of the year. The results of this study show a strong correlation between energy usage and factors such as temperature, time of day and building occupancy. Furthermore, an analysis of the load factors and power factors for each of these residences during different months of the 2017 calendar year revealed possible opportunities for energy saving interventions.**

Keywords—Academic institutions, residential energy usage, load factor, power factor, seasonal energy consumption.

I. INTRODUCTION

Due to a rise in population as well as economic and social development, energy consumption continues to increase worldwide. Urbanisation is a key factor to this, as more than half of the world population live in cities. More than two thirds of the worlds' energy consumption is in cities. As a result of this, cities are responsible for 71% of the greenhouse gas (GHG) emissions in the world. Now more than ever, the planning of urban areas must be and often is influenced by the need to optimise energy usage. To address the increase in energy usage, the amount of energy used, and driver thereof, must be ascertained. Numerous efforts to determine the energy performance characteristics of buildings have been reported [1], [2]. The efforts to improve energy performance in urban areas have resulted in literature, related to approaches such as distributed energy systems, for which, there is a need for more investigation and research [3].

There are many frameworks for measuring the success of energy performance endeavours. The measurement of energy performance entails using a model that meets specified levels of accuracy and relevance. The model represents the usage had there not been an energy saving intervention. The model is compared to measured quantities once the intervention has taken place. The performance measures tend to show discrepancies between the expected saving and the measured savings, or improvements in energy performance.

The energy measuring frameworks aim to integrate the many facets of energy performance in the assessment of energy saving interventions. As facilities and their usage becomes more complex and dynamic, their assessment becomes more tedious [4]. Studies on campuses with complex energy systems have been conducted in [3]. Models for the dynamic measurement of energy performance have been developed and these focus on building or facility applications [5]. Attempts to simplify the development and application of energy performance measurement for management are proposed, where the use of online platforms has emerged as the preferred interface for energy management. Online platforms used on a continuous basis have been shown to resolve discrepancies between expected and measured values to less than 5.5% [6]. Simpler methods of modelling energy performance have been developed because computational cost and complexity of more intricate systems may not be feasible. This is especially relevant when the facility or building is yet to be constructed [7].

The methodology of energy performance assessment for residential energy management, is not as advanced as it is for industrial energy management. This gives rise to the need to establish quantities for energy performance in residential areas of cities [8]. There is a dichotomy that sets residential energy usage apart from industrial energy usage. The consumption behaviour of the residential users varies more than that of industrial users. This is because the residential user has more freedom to use the energy as they wish. Rigorous investigation of urban energy usage shows that the energy users, people in the facilities that are being measured, have an impact on energy performance by how they choose to use energy and manage the appliances which they use [9]. Specific efforts to measure the effect of behaviour on energy performance have been made. Differences in behaviour based on age, gender and standard of living have been considered. Ownership and or responsibility of paying for the energy usage has been considered, which is a part of this study. Policies can be

drawn from the findings of these research activities to reduce the amount of deviations from expected norms. This would result in an improvement of system design and reliability [10],[11] and [12]. According to [11], [12] , [13], the amount of energy saved, once an energy saving intervention is implemented, tends to vary according to the amount of awareness in residential and multi-dwelling buildings.

A new approach to evaluating energy usage in residential areas where user behaviour is a factor thus emerges as a priority for large residential consumers. With energy saving potential ranging from 10- 25% from behavioural changes alone. It is therefore important to evaluate the energy usage of residential users as well as to identify possible drivers to energy usage to further the research development in this field. This study attempts to carry out such an evaluation on 28 residences on the University of Johannesburg Students' Residences.

The rest of the paper is organized as follows: a description of the study area is presented in section 2. The methodology is presented in section 3. The results and discussions are presented in section 4 and the concluding remarks are presented in section 5.

II. Descrption of Study Area

Johannesburg is one of the cities across the globe that is not situated on coast and not on any major river. It is often referred to as the City of Johannesburg (CoJ). It is located in Gauteng province (GP) in the eastern plateau of South Africa (SA) known as the Highveld at an elevation of 1,753 m above sea level. Johannesburg was formerly decentralized into 11 regions but currently, it has been merged and the results of the merger has reduced it into 7 regions. The regions consist of Regions A to G.

The current population of Johannesburg is about 4.4 million which is about 36% of the population of Gauteng and approximately about 8% of the total population of SA. The University of Johannesburg is located in Johannesburg and it is one of the residential universities in SA. It was founded in 2005 following the merger that took place among the Rand Afrikaans University (RAU), the Technikon Witwatersrand (TWR) and the Soweto and East Rand campuses of Vista University. It comprises of four campuses. All of these four campuses are widely spread across the city. The campuses are Auckland Park Kingsway which is the main campus followed by Auckland Park Bunting Road, Doornfontein and Soweto. The staff and students of the university come from more than 50 nations of the African continent and also from other nations across the globe [15], [16], [17].

The university constitutes of 28 residences that are spread over four campuses in the Johannesburg area.These residences housed 6685 students in the 2017 period. Energy consumed in these residences can be mainly attributed to: water heating (geyser and kettle), space heating (heater), cooking(stove), lighting, air cooling(air conditioner, fan) and the use of other electronic equipment.

III. Methodology

Remote data loggers (RDLs) were connected to the meters situated at the student residences. The RDL records data of energy usage at each of these residences. The required data was collected and uploaded from software database known as Metering Online (Belonging to Power Meter Techniques (Energy monitoring and reporting Company)). With the aid of the Metering Online, data for the energy consumptions was collected and analysed for each of the residences. The data for the number of the students who resided at each of the residences was obtained from the management of the University of Johannesburg Student Accommodation and Residence Life. This data was analysed and the results obtained are presented in tabular and graphical form [18], [19], [20].

The aim is to analyse the energy usage of the residences in the University of Johannesburg. The analysis consists of a seasonal and daily profile of all residences. The analysis is carried out with a particular focus on the residences that contribute the most towards electrical energy usage. The analysis also includes the energy usage per student in order to provide an indication of how much energy usage is attributable to each student. The results are presented with a discussion of findings relevant to seasonal and daily usage, load factor and power factor.

IV. Results and Discussions

The results of the data obtained from the analysis carried out using Metering Online are represented in tabular and graphical forms as follows.

A. Seasonal average energy usage

The energy consumption of all 28 residences for the 2017 calendar year is shown in fig.1.

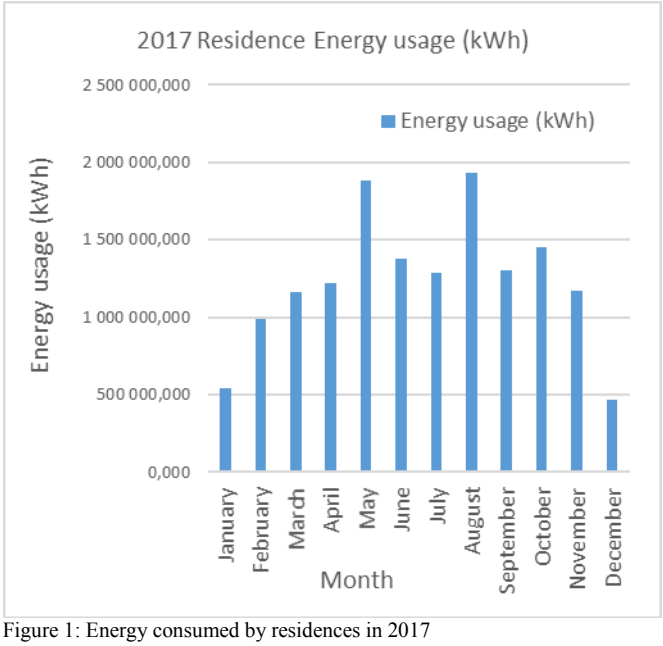


Figure 1: Energy consumed by residences in 2017

When analysing the energy usage trend for the periods January, June, July and December, it is noted that the energy usage shown is lower than what would normally be expected for an energy usage month for the residences. This is due to the lower occupancy during these periods as students are not at the residences for the whole duration of these months due

to recess, resulting in a decrease in occupance. The increase in usage during the colder seasons is evident in Fig.1. 22% of the total campus energy usage is attributable to the residences. When comparing this figure to other academic institutes, it fares well as it is reported that residences generally contribute up to 25% of the campus usage in certain universities [14].

An analysis on the three highest residential consumers of electrical energy was carried out to further investigate the seasonal differences in consumption. The three highest energy using residences where: Sophiatown residence, Kopano residence and Robin Crest residence. The contribution of these residences to the total residential and campus consumption is shown in Tables 1 and 2 respectively.

TABLE 1: RESIDENCE CONTRIBUTION TO TOTAL RESIDENCE ENERGY. USAGE

Residence	Contribution to residential usage (%)
Sophiatown	10,473 %
Kopano	7,329 %
Robin Crest	6,058 %

TABLE 2: RESIDENCE CONTRIBUTION TO TOTAL CAMPUS ENERGY. USAGE

Residence	Contribution to campus usage(%)
Sophiatown	1,85 %
Kopano	1,29 %
Robin Crest	1,07 %

The contribution of residences to the total campus usage is relatively small, thus a small contribution to the total campus usage from the Sophiatown, Kopano and Robin Crest residences is expected.

The seasonal differences in energy consumption are shown in fig.2, where the seasonal consumption of the three main contributing residences are plotted.

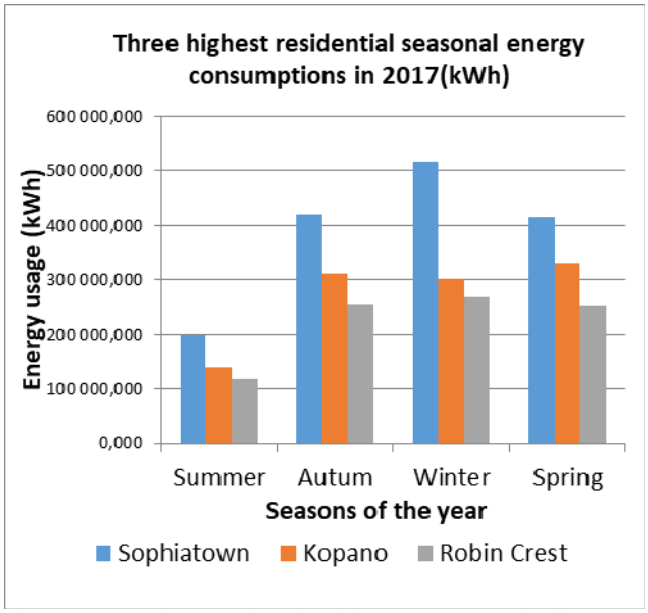


Figure 2: Seasonal consumption of Sophiatown, Kopano and Robin Crest

The highest usage is during the winter season as expected, due the low temperature and high water and space heating requirements. The energy usage of Robin Crest residence is the lowest of the three residences and does not vary as much between the seasons when compared to the energy usage profiles of Sophiatown and Kopano residences. This trend is evident even when the residences are considered on a per student basis. Fig.3 is obtained by dividing the energy consumption per residence consumption by the number of students occupying a given residence. This gives the average amount of electricity attributable to each student on a seasonal basis, and an indication of the average consumption behaviour of each student.

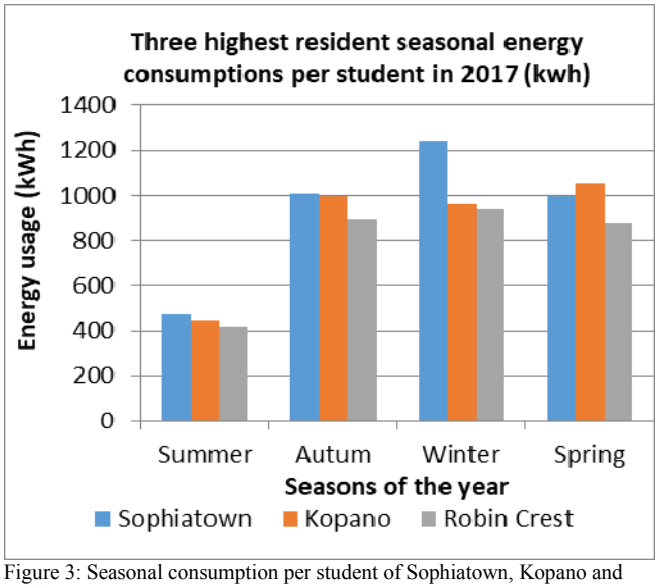


Figure 3: Seasonal consumption per student of Sophiatown, Kopano and Robin Crest

The usage per student for Sophioatown and Kopano residences is higher than that of the Robin Crest residence. Although there are other factors which may contribute to higher energy consumption, such as lighting technology, there is an indication that the students in Robin crest exhibit a lower energy consumption behaviour. This difference would mainly be attributed to the difference in appliances and technologies used in different residences. For instance, Sophia Town Residence uses two heat pumps with back up elements for water heating whilst the other residence may use gas for water heating. Though it is difficult to quantify, human behaviour would also play a significant role in energy usage. When walk through audits were conducted in the residences, it was found that there is a significant amount of energy waste due to the following student activities: bathroom and corridor lights left on permanently, stoves accidentally left running, stoves used for space heating, showers leaking (Loss of hot water, requiring replaced cold water to be re-heated) and appliances in student rooms left on.

B. Load factor

The load factor is defined as the percentage of the peak load, which is equal to the average load in a specified time period. It is a measure of how compatible the installed capacity of the supplied power is to the average power used. The load factor is calculated using the following equation.

$$load\ factor = \frac{100}{TP_{max}} \int_0^T p(t)dt \quad \ldots (1)$$

In (1) the load factor is the time integral of the instantaneous power p(t) over a period of time, T, divided by the period and the maximum power. This is multiplied by 100 to give a percentage.

Load factor is below 70% indicating that there is a high variation in energy usage. High variations in usage are unfavourable as they tend to contribute to higher costs in the form of load factor related tariffs. The load factor of the Sophiatown, Kopano and Robin Crest residences are illustrated in Fig.4 for the 2017 period.

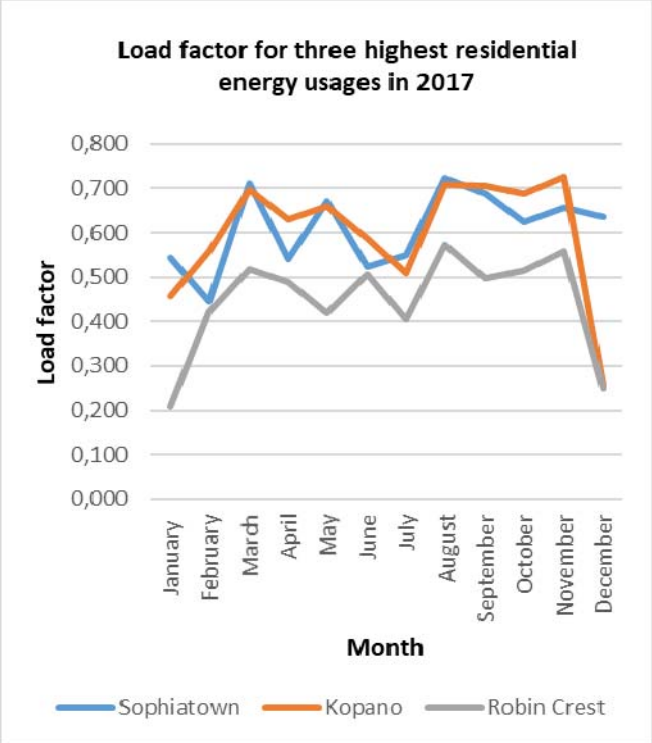


Fig. 4: Monthly average load factor for the Sophiatown, Kopano and Robin Crest residences

The load factor for the Robin Crest was lower than that of the Sophiatown and Kopano residences. The average load factor of all these residences in Fig. 4 was below 70%. This shows that there is potential for improvement in this aspect of energy performance. Particularly, a power factor lower than 70% indicates a potential for reducing the maximum demand required from the utility provider at any given instant. This would subsequently result in a reduction of the demand charge. In order to achieve this, an alternative source needs to supplement the main source during periods where the peak demand is high such that a portion of the required peak demand is drawn from the alternate energy supply. In the case of academic institutions, the alternative energy supply could be backup power in the form of Uninterrupted Power Supplies (UPS) or back up generators.

C. Power factor

Power factor is a well-known energy performance measure/indicator. It is defined as the ratio of active power to apparent power and can be expressed as the percentage of apparent power, which is active power. A low power factor has a direct or implicit cost. It can be directly costly depending on the tariff structure of the facility, or implicitly costly as it reflects the efficiency of energy usage. The monthly power factor for the three highest residential consumers are shown in Fig. 5.

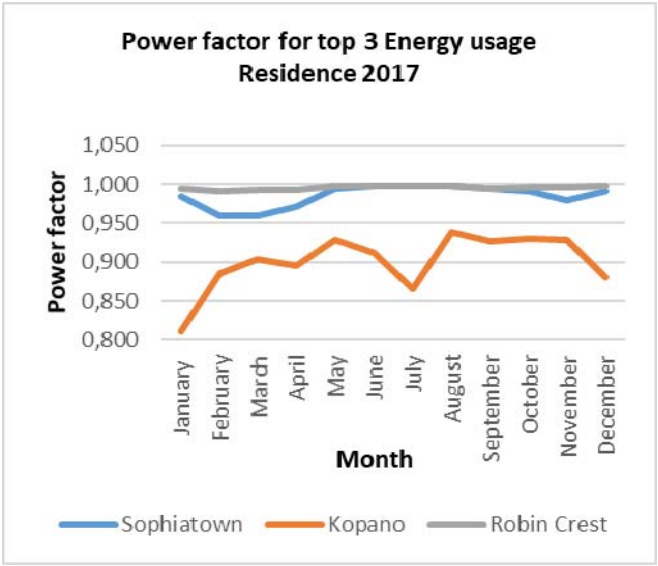


Fig.5: Monthly average power factors for Sophiatown, Kopano and Robin Crest Residences

Each of the residences exhibit relatively high-power factors over 80%, with Kopano residence exhibiting the lowest average power factor (90%) and Sophiatown residence the highest (99%). During months where the power factor is low for residences such as Kopano residence, variable power factor correcting capacitors would be useful in order to increase the power factor and reduce cost to the residence. Further to this, fig.6 shows the daily energy usage of the combined residence profile for one day (14/06/2017 randomly selected), where it is noted that the residences have a high base load (1600kW). It can also be noted that energy is being used during all hours of the day, including periods where students should be sleeping. This could be attributed to energy waste/ leakage.

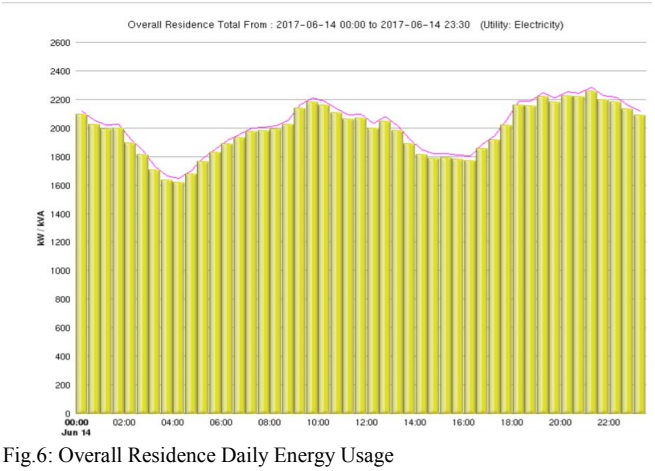


Fig.6: Overall Residence Daily Energy Usage

From this study, a strong correlation can be drawn between energy consumption and temperature/ time of day (energy requirements generally rise with a rise in heating or cooling requirements) occupancy (Energy requirements increase with a increase in residence occupancy and vice versa).

V. Conclusion and Future Recommendations

This study presented an analysis of the energy usage of the twenty-eight residences at the University of Johannesburg during the 2017 calendar year. The results of this study showed a strong correlation between energy usage and factors such as temperature, time of day and building occupancy. Furthermore, an analysis of the load factors and power factors for each of these residences during different months of the 2017 calendar year revealed possible opportunities for energy saving interventions.

It is recommended that the University of Johannesburg put an incentive model in place in such a way that any residence that uses less energy will be rewarded. Also, the actual energy consumption data for all the residences can be presented to students through a web (U-link) as this will be go a long way to cause changes in attitudes most especially when students from one residence notice that their counterparts are using less energy and they are being rewarded for it. Additionally, an energy representative can be appointed in each of the hall of residences who will be educating and motivating students on the need to achieve energy efficiency so as to have a sustainable university.

Acknowledgments

The authors wish to acknowledge the Central Technical Services department of the University of Johannesburg (in Particular Mr Mbambeleli Masala and Mr Gregory James) for the resources provided to complete this project.

References

[1] M. Silva, V. Leal, V. Oliveira and I. M. Horta, "A scenario-based approach for assessing the energy performance of urban development pathways," Sustainable Cities and Society, vol. 40, pp. 372-382, July 2018.

[2] L. Wang, G. Yuan, R. Long and C. Hong, "An urban energy performance and evaluation system and its computer implementation," Journal of Environmental Management, vol. 204, 2017, pp. 684-694.

[3] J. Gao, J. Kang, C. Zhang and W. Gang, "Energy performance and operation characteristics of distributed energy systems with district cooling systems in subtropical areas under different control strategies," Energy, vol. 153, pp. 849-860, June 2018.

[4] M. G. Perroni, S. E. G. da Costa, E. P. de Lima, W. V. da Silva and U. Tortato, "Measuring energy performance: A process-based approach," Applied Energy, vol. 222, pp. 540-553, April 2018.

[5] I. Jaffal and C. Inard, "A metamodel for building energy performance," Energy and Buildings, vol. 151, pp. 501-510, June 2017.

[6] M. Jradi, K. Arendt, F. Sangogboye, C. Mattera, E. Markoska, M. Kjærgaard, C. Veje and B. Jørgensen, "ObepME: An online building energy performance monitoring and evaluation tool to reduce energy performance gaps," Energy & Buildings, vol. 166, pp. 196-209, February 2018.

[7] A. Giretti, M. Vaccarini, M. Casals, M. Macarulla, A. Fuertes and R. Jones, "Reduced-order modeling for energy performance contracting," Energy & Buildings, vol. 167, pp. 216-230, March 2018.

[8] B. M. Marwaha and A. Sharma, "A methodology for energy performance classification of residential building stock of Hamirpur," HBRC Journal, vol. 13, no. 3, pp. 337-352, December 2017.

[9] S. Papadopoulos, B. Bonczak and C. E. Kontokosta, "Pattern recognition in building energy performance over time using energy benchmarking data," Applied Energy, vol. 221, pp. 576-586, July 2018.

[10] S. Pan, Y. Xiong, Y. Han, X. Zhang, L. Xia, S. Wei, J. Wu and M. Han, "A study on influential factors of occupant window-opening behavior in an o□ce building in China," Building and Environment, vol. 133, pp. 41-50, April 2018.

[11] T. Broberg and A. Egüez, "Blame it on the owner- Ownership and energy performance of multi-dwelling buildings," Energy Economics, vol. 72, pp. 108-119, May 2018.

[12] F. Heesen and R. Madlener, "Consumer behavior in energy-efficient homes: The limited merits of energy performance ratings as benchmarks," Energy and Buildings, vol. 172, pp. 405-413, May 2018.

[13] A. Emeakaroha, C. S. Ang and Y. Yan, "Challenges in Improving Energy Efficiency in a University Campus Through the Application of Persuasive Technology and Smart Sensors," Challanges, vol. 1, 2012, pp. 290-318.

[14] BRE, Energy Consumption Guide 54, "Energy Efficiency in Further and Higher Education- Cost-Effective Low Energy Buildings", 1997 BRECSU, Building Research Establishment. [Online].Available: http://www.cibse.org/getmedia/f944f3e0-e047-4c62-af85-43f02533f2de/ECG54-Energy-Use-in-Further-Higher-Education-Buildings.pdf [Accessed: 6 October 2018]

[15] O. O. Ayeleru, F. Ntuli, and C. Mbohwa, "Characterisation of Fruits and Vegetables Wastes in the City of Johannesburg," in Proceedings of the World Congress on Engineering and Computer Science, vol. 2, 2016.

[16] University of Johannesburg, "I care for my community, therefore I choose the University of Johannesburg: 2011 Community Engagement Report," Johannesburg, 2011. [Online].Available:https://talloiresnetwork.tufts.edu/wp-content/uploads/UJ-CE-Brochure.pdf [Accessed: 1 October 2018]

[17] O. O. Ayeleru, F. Ntuli, and C. Mbohwa, "Characterisation of Fruits and Vegetables Wastes in the City of Johannesburg," in Proceedings of the World Congress on Engineering and Computer Science, vol. 2, 2016.

[18] K. Sumer, O. Goktas, A. Hepsag., "The Application of Seasonal Latent Variable in Forecasting Electricity Demand as an Alternative Method", Energy Policy, vol. 37, 2009, pp. 1317-1322.

[19] J. Faraway, " Extending the Linear Model with R," in Statistical heat series. Chapman and Hall/CRC, London, New York, ch. 6, 2006, pp. 126–149.

[20] C. Heylman, Y.G. Kim, J. Wang., "Forecasting Energy Trends and Peak Usage at the University of Virginia", IEEE Systems and information Engineering Design Symposium, vol. 37, 2015, pp. 362-368.

Energy Usage Modelling for Residences of a South African Academic Institution

Sula B. K. Ntsaluba
Dept. of Electrical and Electronic Engineering Technology
University of Johannesburg
Johannesburg, South Africa
ntsaluba.sbk@gmail.com

Esrom M. Malatji
Dept. of Electrical and Electronic Engineering Technology
University of Johannesburg
Johannesburg, South Africa
mahlatsi87@gmail.com

***Abstract*—In residential buildings, various factors often have a significant impact on a building's energy consumption. This paper aims to utilize various existing models to evaluate the sensitivity and influence of each of these factors on the building's energy usage. The factors considered include: average temperature, heating degree days (HDD), cooling degree days (CDD), number of workdays, number of non-workdays and building occupancy. The models considered were: linear two variable and multivariable regression, exponential regression and polynomial regression. The data used for the modeling was that of the energy usage of all the residences of a South African University, during the 2017 academic year. The results of this study revealed that the models developed using polynomial regression produced coefficient of determination (R^2) values ranging from 0.7 to 0.89 in the case of temperature and occupancy, and 0.39-0.69 in the case of workdays and non-workdays, which were the highest model accuracies when compared to those of other models. Analysis of the results also revealed that certain factors such as building occupancy had a greater correlation to the building energy usage. The final model developed (A linear multivariable regression model) achieved an R^2 value of 0.95 indicating the model's high accuracy in predicting the dependent variable (energy consumption) using the factors indicated as independent variables to the model.**

Keywords—average temperature, energy modelling, energy usage, occupancy, regression

I. INTRODUCTION

It is well known that the majority of global electrical energy is derived from diminishing perishable energy sources (fossil fuels). With the ever-increasing cost of energy, the knowledge of energy usage patterns in a residential, commercial and industrial environment has become increasingly important, as this would be the first step to a potential increase in energy conservation through energy saving interventions [1]. Currently, a large degree of effort is placed on the improvement of energy savings through interventions that focus on: equipment/technology replacement, improvement of equipment and operational efficiencies and diversification of in energy sources amongst others. These interventions can be economically unattractive when compared to upfront cost of existing inefficient technologies, leading to a lack of customer acceptance. It is therefore important to explore alternative cost-effective interventions that can be implemented to either directly or indirectly reduce energy consumption. One critical area of focus which has the potential to result in consistent energy savings is the impact on human behaviour, and in this regard, research has shown that not enough emphasis and efforts are placed on the usage awareness and educational aspect of energy usage. According to [2], it is estimated that almost one third of the energy usage in a residential setting can be attributed to waste due to unaware activities of various household appliances. Further research conducted in [3, 4] indicates that with the use of an efficient energy management system and through visualisation of home energy use, a home energy user could potentially reduce their energy usage by between 4-15%.

The building sector in particular consumes and estimated 30–40% of energy globally [5]. Coupled with the global energy crisis, reducing the energy consumption of existing buildings can have a significant impact in overall energy usage reduction. In buildings of academic institutions, energy consumption is mainly driven by factors such as; occupancy, weather conditions, operating hours, type of building and the age of the building [6]. Understanding the energy usage of campuses of academic institutions is an important precondition of understanding how to improve the energy efficiency and make a good energy planning of campus building complexes [7][8].

Energy consumption at the residential sector is grouped into four and this comprises of lighting, heating, cooling and other household appliances [9]. One of the most challenging problems associated with the energy management of residential buildings is the behavioral factors and other conditions mentioned above. In addition, a clear and increasingly recognizable research trend looks at the challenges that impact the energy usage based on activity patterns. In order to improve the energy usage accuracy and confidence levels, the rationale is to provide feedback on usage patterns and derive significant underlying associations between several contextual factors including time of use and user activities. Insights may improve the accuracy and confidence levels of the energy consumption in the residences. This may be used as an additional variable that can enhance electricity forecasting.

This paper aims to model the energy usage of all the residences at a South African University (University of Johannesburg) in order to determine the accuracy and confidence level of the energy usage prediction models applied. The energy data is modeled as a function of parameters such as: average temperature, occupancy,

cooling degree days (CDD), heating degree days (HDD), as well as the number of workdays and non-workdays in a month.

The rest of the paper is structured as follows: in section II presents a description of the study area. The methodology and model formulation equations are presented in section III. In section V the results are presented and analyzed.

II. DESCRIPTION OF THE STUDY AREA

Johannesburg is one of many cities that is situated in the Gauteng Province of South Africa. It is often referred to as the City of Johannesburg (CoJ). It is located in Gauteng province (GP) in the eastern plateau of South Africa (SA). The City is situated 1,753 m above sea level. Johannesburg was formerly decentralized into 11 regions but currently, it has been merged and the results of the merger has reduced it into 7 regions. The regions consist of Regions A to G.

The current population of Johannesburg is about 4.4 million which is almost 36% of the population of Gauteng, and about 8% of the total population of SA. The University of Johannesburg is located in Johannesburg and it is one of the residential universities in SA. It was founded in 2005 following the merger that took place among the Rand Afrikaans University (RAU), the Technikon Witwatersrand (TWR) and the Soweto and East Rand campuses of Vista University. It comprises of four campuses. All of these four campuses are widely spread across the city. The campuses are Auckland Park Kingsway which is the main campus followed by Auckland Park Bunting Road, Doornfontein and Soweto. The staff and students of the university come from more than 50 nations of the African continent and also from other nations across the globe [10], [11], [12]. The load to be analysed is that of 28 residences at UJ. These residences housed 6685 students in the 2017 period.

III. METHODOLOGY AND MODEL FORMULATION

Remote data loggers (RDLs) were connected to the meters situated at the student residences. The RDL records data of energy usage at each of these residences. Data were collected and uploaded into software known as Metering Online [13]. With the aid of the Metering Online, data for the energy consumptions were generated for each of the residences. The name of the company that owns the software is known as Power Meter Technique. The data for the number of the students who resided at each of the residences were obtained from the UJ 2017 Annual report [14]. The data was analysed and the results obtained were represented both in tabular forms and graphically. Other data such as average temperature, HDD and CDD was obtained from accu weather [15], whilst the number of workdays and non-workdays for the measurement period was obtained from the 2017 calendar of the academic institution [16].

Many approaches have been developed in order to model energy data and forecast energy demand. These include: regression, neural networks and support vector methods among others [17][18]. The three regression models that were considered are shown in Table I below.

TABLE I: REGRESSION MODELS

Regression Model	Equation form
Linear two variable	$Y = ax + b$
Linear Multivariable	$Y = ax + by + cz + E$
Exponential	$Y = ae^{bx} + E$
Polynomial 5th order	$Y = ax^4 + bx^3 + cx^2 + dx^1 + E$

All regression models where developed using Microsoft excel. In the case of the polynomial 5th order regression, the 5th order was selected as it provided the highest coefficient of determination value when compared to the lower orders.

IV. RESULTS AND DISCUSSIONS

The results of the data obtained from the analysis carried out using the Microsoft excel data analysis tool are presented below.

A. Monthly energy usage profile (2017)

Fig. 1 shows the monthly energy usage profile for 2017.

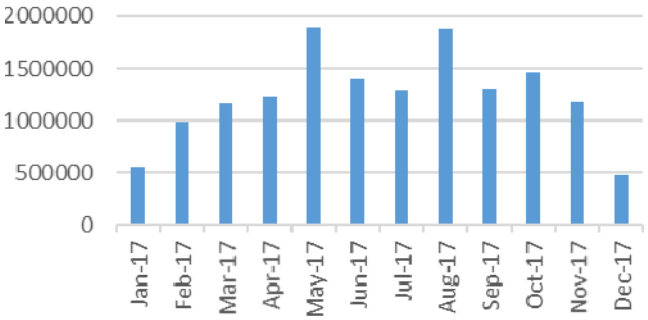


Fig. 1. 2017 Monthly energy usage profile

From Fig. 1, it can be seen that the energy usage profile follows a normal expected trend for residential buildings (reduced energy usage during the summer periods and increased energy usage during the winter periods, as fig.1 indicates), with the exception of certain months (January, June, July and December) where the energy profile is less that what would be expected. This attributed to the reduced occupancy in the residences during these months as these periods coincide with the recess periods of the University.

B. Average Temperature

In order to analyse the impact of average temperature on energy consumption, the linear, exponential and 5th order polynomial regression models were analysed with the average temperature as the independent variable and the energy consumption as the dependent variable. A scatter plot with the measured data as well as the trend graphs for all three regression models is shown in Fig. 2.

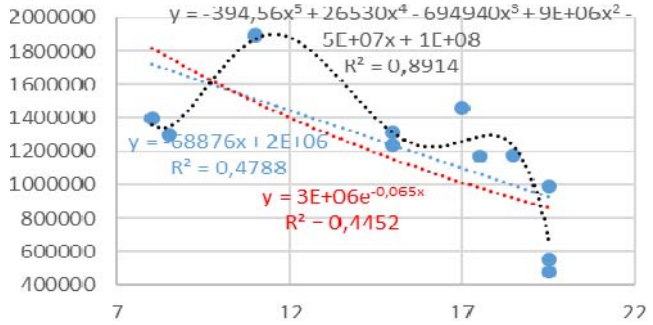


Fig. 2. Average temperature vs energy consumption

Table II highlights the coefficients obtained for the characteristic equations and R^2 values for the three regression models.

TABLE II: TEMPERATURE vs ENERGY CONSUMPTION MODELS

Regression model (Average Temperature)	Equation with Variables	R^2
Linear	$Y = -68876x + 2E+06$	0.47
Exponential	$Y = 3E+06e^{-0,065x}$	0.44
Polynomial 5th order	$Y = -394,56x^5 - 26530x^4 - 694940x^3 + 9E06x^2 - 5E+07x + 1E+08$	0.89

From Fig. 2 and Table II, it can be seen that the model with the highest R^2 value is the 5th order polynomial regression model. The high R^2 highlights that the energy consumed has a significant dependency on the average temperature. The study is concluded in section VI.

C. HDD and CDD

Similarly, the impact of Heating Degree Days (HDD) and Cooling Degree Days (CDD) was analysed. A CCD is defined as the average temperature above a set point temperature above which a building would need to be cooled. Similarly, a HDD is defined as the average temperature below a set point temperature above which a building would need to be heated. A scatter plot with the measured data as well as the trend graphs for all three regression models is shown in Fig. 3 and Fig. 4. while Tables III and IV highlight the coefficients obtained for the characteristic equations and R^2 values for the three regression models.

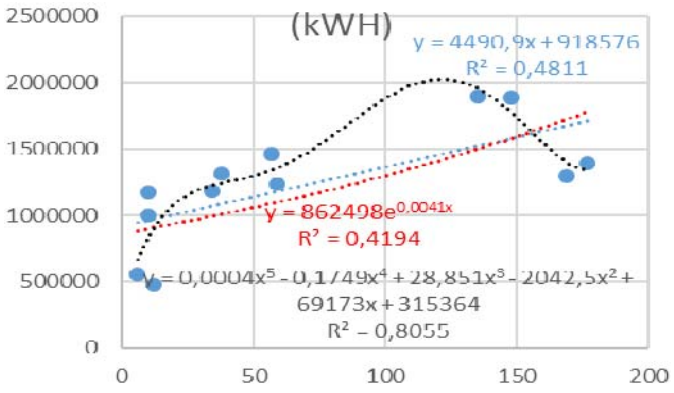


Fig. 3. HDD vs energy consumption

TABLE III: HDD VS ENERGY CONSUMPTION MODELS

Regression model (HDD)	Equation with Variables	R^2
Linear	$Y = 4490,9x + 918576$	0.48
Exponential	$Y = 862498^{e0,0041x}$	0.41
Polynomial 5th order	$Y = 0,0004x^5 - 0,1749x^4 + 28,851x^3 - 2042,5x^2 + 69173x + 315364$	0.801

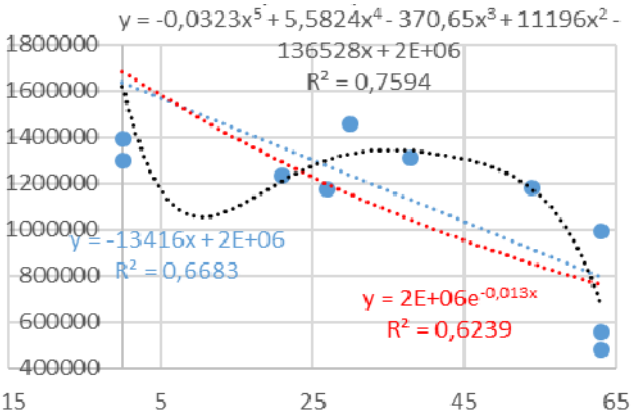


Fig. 4. CDD vs energy consumption

TABLE IV: CDD VS ENERGY CONSUMPTION MODELS

Regression model (CDD)	Equation with Variables	R^2
Linear	$Y = -13416x + 2E+06$	0,6683
Exponential	$Y = 2E+06e^{-0,013x}$	0.62
Polynomial 5th order	$Y = -0,0323x^5 + 5,5824x^4 - 370,65x^3 + 11196x^2 - 136528x + 2E+06$	0.75

From Fig. 3. and Fig. 4. it can be seen that the HDD and CDD show a strong correlation to the buildings' energy consumption. It can therefore be concluded that the buildings energy consumption also has a significant dependency on HDD and CDD. It can also be said that the average temperature may have a stronger correlation as it includes high and low temperatures in one model and has a higher R^2 value specifically for the 5th order polynomial regression model as shown in Tables III and IV.

D. Workdays and Non-workdays

The impact of workdays and non-workdays on the monthly energy consumption was analysed. A scatter plot with the measured data as well as the trend graphs for all three regression models is shown in Fig. 5 and Fig. 6. while Tables V and VI highlight the coefficients obtained for the characteristic equations and R^2 values for the three regression models.

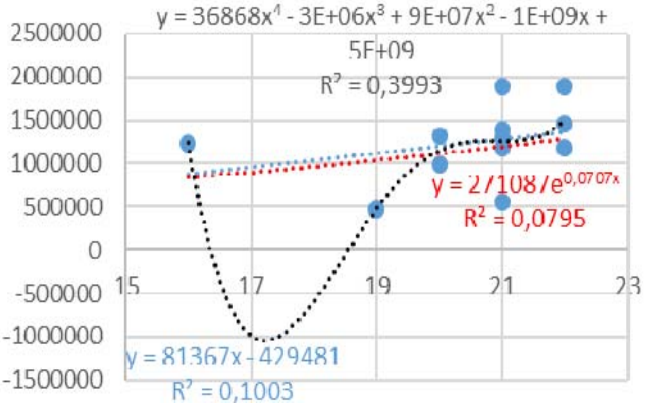


Fig. 5. Workdays vs energy consumption scatter plot and trend lines

TABLE V: WORKDAYS VS ENERGY CONSUMPTION MODELS

Regression model (Workdays)	Equation with Variables	R^2
Linear	$Y = 81367x - 429481$	0,1
Exponential	$Y = 71087e^{0,0707x}$	0.079
Polynomial 5th order	$Y = 36868x^4 - 3E+06x^3 + 9E+07x^2 - 1E+09x + 5E+09$	0.39

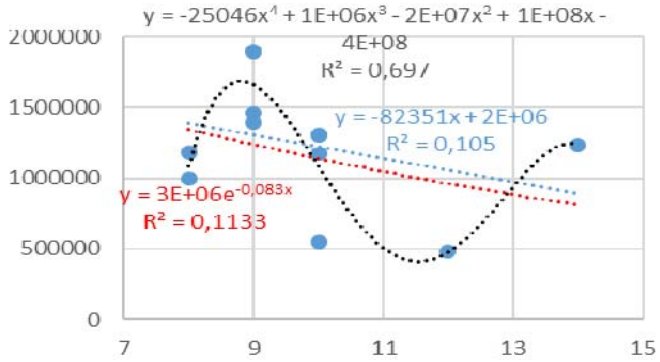


Fig. 6. Non-workdays vs energy consumption

TABLE VI: NON-WORKDAYS VS ENERGY CONSUMPTION MODELS

Regression model (Non-workdays)	Equation with Variables	R^2
Linear	$Y = -82351x + 2E+06$	0,105
Exponential	$Y = 3E+06e^{-0,083x}$	0.113
Polynomial 5th order	$Y = -25046x^4 + 1E+06x^3 - 2E+07x^2 + 1E+08x - 4E+08$	0.69

From Fig. 5. and Fig. 6. it can be seen that the number of work days and non-work days do not show a strong correlation to the buildings' energy consumption (relatively low R^2 values as shown in Table V and VI). It can therefore be concluded that the buildings energy consumption is not heavily sensitive to the type of day, and that there are other factors that are more a significant.

E. Occupancy

In order to analyse the impact of occupancy on energy consumption, the linear, exponential and 5th order polynomial regression models were analysed with the occupancy as the independent variable and the energy consumption as the dependent variable. A scatter plot with the measured data as well as the trend graphs for all three regression models is shown in Fig. 7.

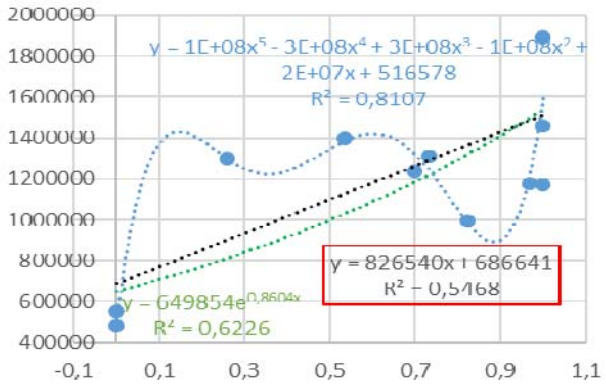


Fig. 7. Occupancy vs energy consumption

Table VII highlights the coefficients obtained for the characteristic equations and R^2 values for the three regression models.

TABLE VII: OCCUPANCY VS ENERGY CONSUMPTION MODELS

Regression model (Occupancy)	Equation with Variables	R^2
Linear	$Y = 826540x + 686641$	0.5
Exponential	$Y = 649854e^{0,8604x}$	0.65
Polynomial 5th order	$Y = 1E+08x^5 - 3E+08x^4 + 3E+08x^3 - 1E+08x^2 + 2E+07x + 516578$	0.81

From Fig. 7 and Table VII, it can be seen that occupancy may have the highest correlation to building energy consumption as the R^2 values for both linear regression and exponential regression were higher than those of all other independent variables. When considering the 5th order polynomial regression model, an R^2 value of 0.81 was achieved which was slightly less than that obtained when using average temperature as an independent variable.

F. Linear Multivariable Regression

The last model considered was a linear multivariable regression model. This model took into account average temperature, CDD, HDD, work days, non-workdays and occupancy as independent variables to predict the Building energy consumption. The results of this model are shown in Table VIII.

TABLE VIII: COEFFICIENTS FOR LINEAR MULTIVARIATE REGRESSION

Independent Variable	Coefficients
Intercept	-955216.18
Average Temperature	-25182.197
HDD	3059.4324
CDD	1888.8431
Working Days	52705.569
Non-Working Days	66168.153
Occupancy	847476.89

From this model it can be seen that an R^2 value of 0.95 was achieved which is the highest of all the models that were considered. Further to this, the P-values were analysed for this model as they would further reveal the significance of each individual factor in predicting the outcome (building energy usage). The results are shown in Table 9.

TABLE IX: P-VALUES

Independent Variable	P-value
Intercept	0.733014
Average Temperature	0.718369
HDD	0.375981
CDD	0.871305
Work Days	0.524848
Non-working Days	0.544810
Occupancy	0.035437

From the results in Table IX, it can be seen that occupancy has the lowest P-Value, indicating the highest significance in predicting the outcome. When comparing the overall results of this paper, it can be noted that linear multivariable regression has provided the most accurate modelling with the highest R^2 value. This conclusion is consisted with that of work conducted in [19] and [20].

V. Conclusion and Future Work

This research presented in this paper evaluated the various factors influencing building energy usage for residences of an academic institution in South Africa. The results showed that the 5th order polynomial regression model yielded the most accurate prediction results when considering only one independent variable. The results also indicated that occupancy showed the highest significance in predicting the output. This was further verified with a linear multivariable regression model which yielded similar results. As a future consideration, additional factors influencing energy usage that have been indicated in other studies will be considered. Benchmarking of results with those of similar papers will also be considered. Finally, models that will accurately forecast future energy usage with certain given inputs will be considered as this is a practical need that can be applicable in assisting facility managers.

Acknowledgments

The authors wish to Ms Maria Michael for her assistance in reviewing and structuring the research content of this paper. The authors also wish to acknowledge the University of Johannesburg for supplying the relevant data required to complete the research in this paper.

References

[1] R. L. Sawyer, J. M. Anderson, E. L. Foulks, J. O. Troxler and R. W. Cox, "Creating low-cost energy-management systems for homes using non-intrusive energy monitoring devices," IEEE Energy Conversion Congress and Exposition, San Jose, pp. 3239-3246, 2009.

[2] Abubakar, S. Khalid, M. Mustafa, H. Shareef, and M. Mustapha, "Application of load monitoring in appliances's energy management-a review," Renewable and Sustainable Energy Reviews, vol. 67, pp. 235-245, 2017.

[3] M. Coaster. "Smart Metering Implementation program, In-home display units," Supporting documents, 2008. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/42731/233-smart-metering-imp-in-home.pdf [Accessed: 5 October 2018].

[4] K. Ehrhardt-Martinez, K. Donnelly, and J. Laitner, "Advanced metering initiatives and residential feedback programs: A meta-review for household electricity-saving opportunities," American Council for an Energy-Efficient Economy: Washington, DC, USA, 2010. [Online]. Available: https://aceee.org/sites/default/files/publications/researchreports/e105.pdf [Accessed: 5 October 2018]

[5] J. Woo, C. Menassa, Virtual retrofit model for aging commercial buildings in a smart grid environment, Energy Build., vol. 80, pp.424-435, 2014.

[6] P. Govender, 'Energy audit of the Howard College Campus of the University of KwaZulu-Natal', University of KwaZula-Natal, 2005. [Online]. Available: https://researchspace.ukzn.ac.za/handle/10413/5087 [Accessed: 5 October 2018]

[7] K. Brown, M. Anderson, J. Harris, Setting enhanced performance targets for a new university campus: benchmarks vs. energy standards as a reference in: Proceedings of the 2002 ACEEE Summer Study of Energy Efficiency in Buildings, American Council for an Energy-Efficient Economy, Washington, pp. 29–40, 2002.

[8] K. I. Evangelinos, N. Jones, E.M. Panoriou, "Challenges and opportunities for sustainability in regional universities: a case study in Mytilene Greece", J. Clean. Prod., vol 12, pp.1154–1161, 2009.

[9] E. Hittinger, K. A. Mullins, and I. L. Azevedo, "Electricity consumption and energy savings potential of video game consoles in the United States," Energy Efficiency, vol. 5, pp. 531-545, 2012.

[10] O. O. Ayeleru, F. Ntuli, and C. Mbohwa, "Municipal Solid Waste Composition Determination in the City of Johannesburg," in Proceedings of the World Congress on Engineering and Computer Science, WCECS 2016 - San Francisco, United States, pp.625-629, 2016.

[11] University of Johannesburg, "I care for my community; therefore I choose the University of Johannesburg: 2011 Community Engagement Report," Johannesburg, 2011. [Online]. Available: https://talloiresnetwork.tufts.edu/wp-content/uploads/UJ-CE-Brochure.pdf [Accessed: 1 October 2018]

[12] O. O. Ayeleru, F. Ntuli, and C. Mbohwa, "Characterisation of Fruits and Vegetables Wastes in the City of Johannesburg," in Proceedings of the World Congress on Engineering and Computer Science, Vol.2, 2016.

[13] Power Meter Techniques, '2017 UJ residences monthly energy consumption', 2018. [Online]. Available:http://www.meteringonline.com/admin/index.php [Accessed: 1 October 2018]

[14] Accuweather, 2017 'Temperature data, Johannesburg, South Africa', 2018 [Online]. Available: http://www.accuweather.com/ [Accessed: 1 October 2018]

[15] University of Johannesburg, 2018 'University of Johannesburg Annual report 2017, South Africa', 2018. [Online]. Available: https://www.uj.ac.za/about/PublishingImages/Pages/Annual-Reports/UJ-Annual-Report-2017-DHET.pdf [Accessed: 3 October 2018]

[16] University of Johannesburg, 2017 'University of Johannesburg Acedemic calendar 2017', 2018. [Online]. Available: https://www.uj.ac.za/studyatUJ/PublishingImages/Pages/DFC-Campus-Timetables/E-UJ%20ACADEMIC%20CALENDAR%202017.pdf [Accessed: 3 October 2018]

[17] H. Hahn et al., "Electric Load Forecasting Methods: Tools for Decision Making", European Journal of Operational Research, vol. 199, no. 3, pp. 902-907, 2009.

[18] K. Sumer, O. Goktas, A. Hepsag., "The Application of Seasonal Latent Variable in Forecasting Electricity Demand as an Alternative Method", Energy Policy, vol. 37, no. 4, pp. 1317-1322, 2009.

[19] J. Faraway, "Introduction" in Extending the Linear Model with R. Boca Raton: Chapman & Hall/CRC, 2006.

[20] J. Faraway, "Extending the Linear Model with R," in Statistical heat series. Chapman and Hall/CRC, London, New York, 2006, ch. 6, pp. 126-149.

[21] C. Heylman, Y.G. Kim, J. Wang., "Forecasting Energy Trends and Peak Usage at the University of Virginia", IEEE Systems and information Engineering Design Symposium, vol. 37, no. 1, pp. 362-368, 2015.

Author Index