2023 International Conference on Artificial Intelligence and its Applications
https://doi.org/10.59200/ICARTI.2023.004

Reflections on Feature Engineering and Design Using Causal Machine Learning (CML) for African Swine Fever (ASF) Diagnosis

Steven Lububu and Dr. Boniface Kabaso.
Information Technology
Cape Penisula University of Technology
Cape Town, South Africa
stevenlububu23@gmail.com, KabasoB@cput.ac.za

***Abstract*— Feature engineering is a crucial step in the process of machine learning, where raw data is transformed into meaningful features that can effectively represent the underlying patterns and relationships in the data. The goal is to improve the performance of machine learning models by providing them with more informative and meaningful input features. Automated feature engineering techniques, such as genetic algorithms, can also be used to automatically generate and optimise features. These methods search a space of potential features and select or create features based on their impact on the model's performance. Overall, feature engineering plays a crucial role in machine learning by enabling models to exploit the most relevant and informative aspects of the data, thereby improving their accuracy, robustness, and interpretability. This paper reports empirical studies aimed at demonstrating which types of technical features are best suited to establish relationships between ASF viruses and clinical symptoms to accurately diagnose ASF disease. Various machine learning models such as neural networks, decision trees, random forests, linear regression, and Bayesian regression accept ASF features and provide predictions. The experiment demonstrates the extent to which the machine learning model can establish correlations between ASF viruses and clinical symptoms by independently analysing the required feature. The focus is on establishing relationships between ASF viruses and clinical symptoms for diagnosis. Data from the European Union Reference Laboratory for African swine fever (ASF) was collected for the study. This paper provides essential information on ASF datasets based on the interpretation of results obtained by using appropriate samples and validated tests in combination with information from laboratory tests on ASF disease epidemiology, scenario, clinical signs, and lesions caused by different virulence. The study proposes to use causal ML to establish relationships between ASF viruses and symptoms to improve the accuracy of the ASF disease. In this study, the performance and validation of the models were measured using metrics such as R-squared, mean absolute error (MAE) and mean square error (MSE).**

Keywords*— *Feature Engineering, Causal Machine Learning (CML), and Accuracy.

I. INTRODUCTION

The performance of machine learning depends heavily on the representation of the feature engineering. For this reason, data scientists spend much of their work developing preprocessing pipelines and data transformations [1]. Feature engineering is the process of selecting, transforming, and creating new features (input variables) from existing data to improve the performance of a machine learning model. It involves identifying and extracting relevant information from raw data that can be used effectively to represent the problem at hand [1].

Feature engineering is critical because the quality and relevance of the features used as input to a machine learning algorithm significantly affect the model's ability to learn, and the accuracy of its predictions. By transforming or creating new features, feature engineering aims to improve the representation of the data, capture important patterns or relationships, and remove noise or irrelevant information [2]. Feature engineering has been the focus of interest for some time and is still limited or insufficiently explored. Therefore, more determined attempts are needed to advance the process of feature engineering in the context of learning algorithms to predict better outcomes and behaviours [3],[4]. With the huge amounts of data available and the resulting demands on artificial intelligence and good machine learning techniques, new problems arise and new approaches to feature engineering are needed [5].

To apply feature engineering, the model must preprocess its input data by adding new features based on the other features [6]. These new features can be ratios, differences, or other mathematical transformations of existing features, similar to the equations that human analysts design. They construct new traits such as body mass index, wind chill or the ratio between triglycerides and HDL cholesterol to better understand the interactions between existing traits [2],[7].
Some common techniques used in feature engineering are:

Feature extraction: this involves deriving new features from existing features. For example, extracting the day of the week or time from a timestamp, or calculating statistical measures such as mean, median or standard deviation from a set of values.

Feature transformation: This involves applying mathematical transformations to the features to make them more suitable for the learning algorithm. Examples include scaling features to a specific range, applying

logarithmic or exponential transformations, or using mathematical functions such as square roots.

One-hot-encoding: This technique is used to convert categorical variables into binary features that can be easily understood by machine learning algorithms. Each category is represented by a binary value (0 or 1) indicating the presence or absence of that category.

Feature selection: This involves selecting a subset of the most relevant features from the available set. This helps to reduce dimensionality, eliminate noisy or redundant features, and improve the efficiency and interpretability of the model.

Domain-specific feature engineering: In some cases, expertise can be used to develop features that are specific to the problem at hand. For example, in natural language processing, features such as word frequency, n-grammes or sentiment values can be developed to effectively capture textual information.

Overall, feature development is an iterative process that requires deep understanding of the problem domain, data exploration and experimentation. It plays a crucial role in improving the performance and generalisation capabilities of machine learning models. This paper provides a comprehensive overview of feature engineering methods and techniques for improving model accuracy in unseen ASF data. It also presents applications of feature engineering in text classification and clinical text classification to achieve high performance of predictive learning algorithms in terms of model accuracy.

II. BACKGROUND AND RELATED WORK

Feature engineering is the transformation of raw data into a set of meaningful features that can be used as input to machine learning algorithms. The purpose of feature engineering is to improve the performance of machine learning models by creating informative, discriminative, and relevant representations of the data.[8],[9],[10]. A good feature is a big problem in learning predictive models because it can largely determine the performance of the underlying problem. A good feature should 1) be informative, 2) not change under a series of transformations, and 3) be fast to compute. Data analysts often start by examining the existing features of the underlying problem sing their own knowledge domain to find suitable features [11],[12],[13].

Feature engineering is the basis for learning algorithms. It is the process of using expertise about the data to create features that make the learning algorithm functional [14],[15]. To use predictive learning algorithms for underlying problems, the inputs must be converted into such a format that the algorithm can understand them and provide the exact class to which an entity belongs, as well as future predictions [16],[17],[18]. Figure 1 shows the general framework of feature engineering used in this work.

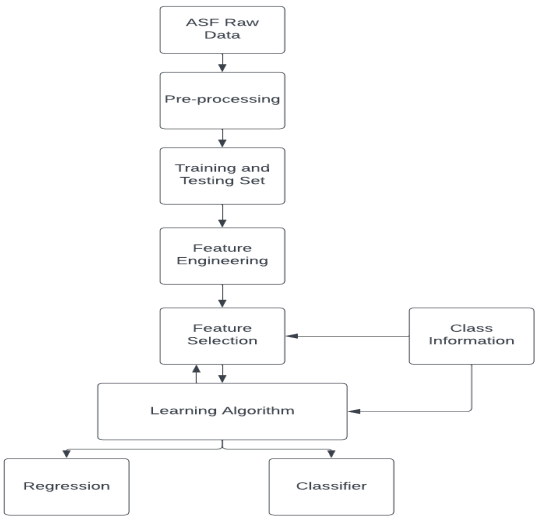

FIGURE1: Framework of Feature Engineering for ASF diagnosis

ASF's feature engineering takes place after data cleaning and preparation before the model is trained and tested. The main goal is to provide the predictive learning algorithm with a better representation of the data.

STEP 1: COLLECTION OF RAW DATA

Data collection took place at the European Union Reference Laboratory for African swine fever (ASF) [19]. The aim of this work is to determine the relationship between ASF viruses and clinical symptoms by applying causal analysis ML. The data collected were text data.

Table 1 shows the details of the data collected by the laboratory for ASF. This table (Table I) provides essential information on ASF datasets based on the interpretation of results obtained from the use of appropriate samples and validated tests in combination with the information from the laboratory tests on ASF disease epidemiology, scenario, clinical signs, and lesions caused by different virulence.

TABLE I: COLLECTION OF ASF FEATURES

CAUSES	EFFECTS				
INFECTIOUS VIRUS OF ASF.	**Observed Clinical Signs**	**Lesions**	**Temperature**	**Mortality**	**ASF Clinical Form**
Hemagglutinating Encephalomyelitis Virus (HEV)	Anorexia, Inactivity hyperpnoea, Cutaneous hyperaemia	Nasal cavity, Tonsils, Vomiting, Fever	41-42°C	100 % [1-4 days]	**Peracute ASF:** Highly virulent
Porcine Reproductive and Respiratory Syndrome virus (PRRSV).	Usually, pigs die suddenly without clinical signs.	Nasal cavity, Tonsils, Vomiting, Fever	41-42°C	100 % [1-4 days]	**Peracute ASF:** Highly virulent
Mycoplasma hyopneumoniae	Fever	No lesions are evident in organs.	41-42°C	100 % [1-4 days]	**Peracute ASF:** Highly virulent
Salmonella Choleraesuis	Lethargy	Leg weakness, discoloration of the legs, ears, tails, fever.	40-42°C	90-100 % [6-9 days]	**Peracute ASF:** Highly virulent
Salmonella Typhimurium	Lethargy	The area around the tail may be soiled with bloody faeces.	40-42°C	90-100 % [11-15 days]	**Peracute ASF:** Highly virulent
Rotavirus and Transmissible	Constipation or diarrhea.	Running stomach	40-42°C	90-100 % [6-9 days]	**Peracute ASF:** Highly virulent
Gastroenteritis (TGE) virus, and Isosporasuis.	Constipation or diarrhea.	Running stomach	40-42°C	90-100 % [11-15 days]	**Peracute ASF:** Highly virulent
RNA virus	Haemorrhagic splenomegaly	Bleeding, red blood cells, Haemorrhages under the skin, bloody from the nose/mouth and a discharge from the eyes.	40-42°C	90-100 % [6-9 days]	**Peracute ASF:** Highly virulent
Mycobacterium species, often M avium	Haemorrhagic lymphadenitis	Enlarged neck, Small and enlarged Intestines, Clotted blood in the stomach.	40-42°C	90-100 % [6-9 days] or [11-15 days]	**Peracute ASF:** Highly virulent
Erysipela	Petecchial haemorrhages	Respiratory problem.	40-42°C	90-100 % [6-9 days]	**Peracute ASF:** Highly virulent
The Porcine Reproductive and Respiratory Syndrome (PRRS) and Streptococcal infections	Petecchial haemorrhages.	Respiratory, Reproduction, Pneumonia, and abortion problems.	40-42°C	90-100 % [6-9 days] or [11-15 days]	**Peracute ASF:** Highly virulent
Fowl Adenovirus serotype 4 (FAV-4)	Hydropericardium syndrome (HPS)	Liver congestion, Ascites with yellowish fluid in kidney and Liver, Hepatomegaly	40-42°C	90-100 % [6-9 days]	**Acute ASF:** Highly virulent
Specific Pathotypes of Escherichia coli	Perirenal oedema	Facial and Body swelling	40-42°C	90-100 % [6-9 days]	**Acute ASF:** Highly virulent
Similar to those observed in the acute form.	Less fever, Depression, Loss of appetite, Painful walking;	Perirenal edema, Partial hyperemic, splenomegaly with focal infarction, Dark red hematomas.	?	30-70% [7-20 days]	**Subacute:** Moderately virulent.
Similar to those observed in the acute form.	Slight fever, Mild respiratory, Moderate-joint swelling.	Edematous lymph nodes, fibrinous pericarditis;	40-40. 5 °C	30% [1month]	**Chronic ASF:** Low virulent
Swine pox	Red spots	Red spots, Circular red lesions on the flank, Circular red lesions on abdomen, Circular red lesions on face and head	40-40. 5 °C	30% [1month]	**Chronic ASF:** Low virulent

STEP 2: PRE-PROCESS THE DATA

Format raw data by aligning, unifying, grouping, and cropping. Removing noisy, dirty data (missing, duplicate, ill-formed, wrong values, etc.) by pairwise or listwise deletion, by calculating imputation (mean substitution, regression), by stochastic simulation, principle of least damage, sampling error, population parameters, dispersion, statistical power, etc., as approaches vary widely. The correctness of the data influences the model accuracy. The model is trained based on the correctness of the data.

STEP 3: FEATURE ENGINEERING (CONVERSION OF FEATURES AND CREATION OF ADDITIONAL FEATURES)

The cleaned data is still unprocessed and much of the data is unusable, so the data is filtered, dissected, and converted to create features for modelling. Feature creation is a difficult task that requires analysis and expertise. Some common methods of feature engineering are PCA, kernel PCA, partial least squares, discretisation, information entropy theory, ICA, MDA, latent factors, statistical moments, mutual information theory, generalised least squares, noise reduction, spot extraction, autocoding, edge detection, weighing, smooching, etc.

The raw data collected was in categorical (text) form, which was not suitable for machine learning processing. This part of the study took a lot of time to convert the categories into numbers. After data pre-processing, the pre-processed data was filtered using the data extraction process. The selected data was then used to create a learning model. In addition, the correlation matrix was used to check the correlation between the variables. This is discussed in the next section Findings and Results.

After data pre-processing, the pre-processed data were filtered using the data extraction process. The selected categorical features were then converted into numbers and used to create learning models.

A. TURNING THE CATEGORICAL FEATURES INTO NUMBERS

The following figure (Fig. 2) shows the conversion of categorical characteristics into numerical values.

```
In [11]:
#Turning categorical features into numbers
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import ColumnTransformer


categorical_features = ["Viruses","Signs"]
one_hot = OneHotEncoder()
transformer = ColumnTransformer([("one_hot",
                                  one_hot,
                                  categorical_features)],
                                  remainder="passthrough")
transformed_x= transformer.fit_transform(df)

In [14]:
transformed_x

Out[14]: <18x22 sparse matrix of type '<class 'numpy.float64'>'
        with 36 stored elements in Compressed Sparse Row format>
```

FIGURE 2: Conversion of Categorical into Numerical Features

B. TRANSFORMING ASF DATASET (VIRUSES AND SIGNS) INTO 0s AND 1s

The next figure (Fig. 3) shows the conversion of the ASF dataset into 0s and 1s.

```
In [15]:
dummies = pd.get_dummies(df[["Viruses"]])
dummies
```

Out[15]:

	Viruses_Choleraesuis	Viruses_Erysipela	Viruses_Escherichia	Viruses_FAV	Viruses_HEV	Viruses_Hyopneumoniae	Viruses_Mavium	Viruses_PRRS	Viruses_PRI
0	0	0	0	0	1	0	0	0	
1	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	1	0	0	
3	1	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	
5	0	0	0	0	0	0	0	0	
6	0	0	0	0	0	0	0	0	
7	0	0	0	0	0	0	0	0	
8	0	0	0	0	0	0	1	0	
9	0	1	0	0	0	0	0	0	
10	0	0	0	0	0	0	0	1	
11	0	0	0	1	0	0	0	0	
12	0	0	1	0	0	0	0	0	
13	0	0	0	1	0	0	0	0	
14	0	0	1	0	0	0	0	0	
15	0	0	0	1	0	0	0	0	
16	0	0	1	0	0	0	0	0	
17	0	0	0	0	0	0	0	0	

```
In [16]: 1 dummies = pd.get_dummies(df[["Signs"]])
         2 dummies
```

Out[16]:

	Signs_Anorexia	Signs_HPS	Signs_Haemorrhages	Signs_Haemorrhagic	Signs_LessFever	Signs_Lethargy	Signs_Perirenal	Signs_RedSpots
0	1	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0
3	0	0	0	0	0	1	0	0
4	0	0	0	0	0	1	0	0
5	1	0	0	0	0	0	0	0
6	1	0	0	0	0	0	0	0
7	0	0	0	1	0	0	0	0
8	0	0	0	1	0	0	0	0
9	0	0	1	0	0	0	0	0
10	0	0	1	0	0	0	0	0
11	0	1	0	0	0	0	0	0
12	0	0	0	0	0	0	1	0
13	0	0	0	0	1	0	0	0
14	0	0	0	0	1	0	0	0
15	0	0	0	0	1	0	0	0
16	0	0	0	0	1	0	0	0
17	0	0	0	0	0	0	0	1

FIGURE 3: Transformation of ASF Dataset into 0s and 1s.

STEP 4: SELECTION OF FEATURES

Feature selection is a crucial step in the machine learning pipeline, as it involves selecting the most relevant and informative features (input variables) for a predictive model. The goal of feature selection is to improve model performance, reduce overfitting and speed up training and inference. The choice of method should depend on the problem and the data set [20], [21].

The feature selection process usually involves the following steps:

1) Preparing the dataset: Start by collecting or preparing the dataset that contains both the input characteristics (independent variables) and the target variable (dependent variable) that you want to predict.

2) Importance of the characteristics: Assess the relevance and importance of each feature in the dataset. This can be done using various statistical methods or machine learning algorithms.

3) Model-based methods: These methods use a machine learning algorithm to train a model on the entire dataset and then evaluate the importance of each feature based on its contribution to the model's performance. Examples include decision trees, random forests, and gradient boosting algorithms. The model-based methods were used to train the CML model on the entire ASF dataset and then evaluate the importance of each feature based on its contribution to the model's performance.

4) Evaluation and validation: After selecting the subset of features, it's important to evaluate the performance of the model using cross-validation or a separate validation set. This step ensures that the feature selection has indeed improved the generalisation ability of the model.

STEP 5: MODELLING AND PERFORMANCE MEASUREMENT

Build models and iteratively use the performance of the learning algorithms to evaluate the quality of the selected features, e.g., wrapper models, cross-validation, etc. This research focused on regression models rather than classification models. This was done to avoid confusion with the classified and tested ASF datasets from the laboratory. Nevertheless, a classification model was also tested. On the other hand, an attempt was made to test the data processing systems using regression models. The following algorithms such as multiple Linear regression, Classification, ANN, Decision Trees, and Bayesian regression were selected for testing. The ML model selection of the features for each algorithm was based on their performance. For this study, performance and validation were measured using metrics such as R-squared, mean absolute error (MAE) and mean square error (MSE). The machine learning models were selected in the assessment based on several matrices that were discussed individually. R^2, MAE and MSE are commonly used in statistical and machine learning models to evaluate the performance of the model.

In the following, the individual metrics are explained together with the corresponding formulae:

- R^2 (coefficient of determination):
  The R-squared measures the proportion of variance in the dependent variable that can be explained by the independent variables in a regression model. It indicates how well the model fits the data.

  Formula:
  R-squared = 1 - (SSR/SST)
  where SSR is the sum of squared residuals (the sum of squared differences between the predicted values and the actual values) and SST is the total sum of squares (the sum of squared differences between the actual values and the mean of the dependent variable).
- MAE:
  MAE stands for the average absolute difference between the predicted and the actual values. It is a measure of the average size of the errors caused by the model.

  Formula:
  $MAE = (1/n) * \Sigma|yi - xi|$

where n is the number of data points, yi is the predicted value and xi is the corresponding actual value.

- MSE:
  The MSE calculates the average of the squared differences between the predicted values and the actual values. It penalises larger errors more than MAE.
  Formula:
  MSE = (1/n) * Σ (yi - xi)^2 where n is the number of data points, yi is the predicted value and xi is the corresponding actual value.

In the above formulae, Σ stands for the sum symbol, and yi and xi are paired values of predicted and actual values. These metrics provide quantitative measures for assessing the performance and accuracy of regression models, with R^2 indicating goodness of fit and MAE and MSE quantifying average errors. The results obtained using model-based methods are explained in detail in the Results and Discussions section.

III. EXPERIMENT DESIGN AND METHODOLOGY

Different machine learning models are capable, to varying degrees, of synthesising different kinds of mathematical expressions. If the model can learn to analyse a constructed feature itself, there was no reason to construct the feature in the first place. Empirical evidence of a model's ability to analyse a particular kind of expression shows whether constructed features of that kind could be useful to the model. To investigate these relationships, we used ASF datasets containing the inputs and outputs corresponding to a particular type of constructed feature. If the machine learning model can learn to reproduce this feature with a small error, it means that the model could have learned this feature without help.

For this study, only machine learning regression models were considered. However, a random forest classifier was also tested for comparison purposes. We selected the following machine learning models based on their relative popularity and their different approaches: Linear regression, Random Forest Classifier, ANN, Decision Trees, and Bayesian regression. To determine the relationships between the ASF viruses and the clinical symptoms of some of these machine learning models, each experiment was run several times and the result of the best run was used for comparison. These experiments were conducted in the Python programming language using the following third-party packages: Scikit-Learn [9] and TensorFlow [10]. The Python source code for these experiments is available on the author's GitHub page.

A. LINEAR REGRESSION MODEL

The R^2, MAE and MSE were used to evaluate this model. In addition, the following conditions were applied to evaluate the linear regression model:

i) Linearity:
The relationship between the variables Viren (Independent) and Sign (Dependent) should be linear. This was tested using the residual curve.

ii) Near normal residuals:
The residuals were normally distributed and had their midpoint at zero. The presence of unusual observations (noise) is not allowed under this condition.

iii) Constant variables:
This condition was checked against the residual plot.

iv) Significance (probability value):
The probability value (p-value) determines the overall significance of the model. If the p-value is less than 0.05, the model can be considered significant.

v) R-squared error:

The R-squared error, $R2$ of a regression model is:

$$R^2 = 1 - \frac{\sum_{n=1}^{N}(y_n - \hat{y}_n)^2}{\sum_{n=1}^{N}(y_n - \bar{y})^2}$$

Explanation of the Equation (1):

$\hat{y}n$ is the nth predicted value, $\bar{y}$ is the mean of the response values. The closer the value is to 1, the more the points fall on the regression line, so the stronger the linear relationship between the two sequences. $R2 = 1$ means that the two sequences have a perfect linear relationship, while $R2 = 0$ means that they have no linear relationship at all. $R2$ is a measure of the goodness of the model. The larger $R2$ is, the better the model is. According to [22], a model is very good if $R2 = 1$, and very poor if $R2 = 0$. If the model fits very poorly, a negative R-squared value may result. However, $R2$ alone cannot be used to assess model performance, as over-fitted models can yield larger $R2$ values.

i) Adjusted R-squared:

$$R^2_{adj} = 1 - \left[\left(\frac{N-1}{N-k-1}\right)(1-R^2)\right]$$

Explanation of the Equation (2):
N is the number of observations(signs), k is the number of independent variables (viruses)

ii) Residual Standard Error ($\hat{\sigma}$):

This equation is calculated from the sum of the squared errors.

$$\hat{\sigma} = \sqrt{\frac{SSE}{DF}}$$

Explanation of the Equation (2):
This model would give an average $\pm\hat{\sigma}$ error. If we assume that the residuals are assigned, then $\hat{\sigma}$ can be used to determine 2/3 or 65 per cent of the outcome in the $\pm\hat{\sigma}$ range, and 95% of the prediction would be in the $\pm 2\hat{\sigma}$ range.

B. RANDOM FOREST CLASSIFIER MODEL

Classification functions were performed to determine the posterior distribution of viruses (independent variables) and features (dependent variables). R^2, MAE and MSE were used to evaluate this model. R^2 refers to "(1)," MAE and MSE refer to "(2)".

C. ARTIFICIAL NEURAL NETWORK (ANN) MODEL

The architectures of ANN were tested. The performance of ANN was evaluated using accuracy, R^2 refers to "(1)," MAE and MSE refer to "(2)".

D. BAYESIAN REGRESSION (BR) MODEL

The BR was used to determine the posterior distribution of viruses (independent variables) and characteristics (dependent variables). The R^2, MAE and MSE were used to evaluate this model. R^2 refers to "(1)," MAE and MSE refer to "(2)".

E. DECISION TREES (DT) MODEL

The DT was administered to determine the posterior distribution of viruses (independent variables) and traits (signs). The R^2, MAE and MSE were used to evaluate this model. R^2 refers to "(1)," MAE and MSE refer to "(2)".

IV. RESULTS AND DISCUSSIONS

A. CORRELATION MATRIX BETWEEN VIRUSES AND SIGNS

Figure 4 presents the correlation matrix between the collected features. In this correlation matrix, each virus corresponds to its appropriate sign. The correlation matrix establishes relationships between viruses and signs. The figure (Fig. 4) shows the correlation matrix of base features and dependent variables (Viruses and Signs).

Out[26]:

	Viruses_Choleraesuis	Viruses_Erysipela	Viruses_Escherichia	Viruses_FAV	Viruses_HEV	Viruses_Hyopneumoniae	Viruses_Mavium	V
Viruses_Choleraesuis	1.000000	-0.058824	-0.108465	-0.108465	-0.058824	-0.058824	-0.058824	
Viruses_Erysipela	-0.058824	1.000000	-0.108465	-0.108465	-0.058824	-0.058824	-0.058824	
Viruses_Escherichia	-0.108465	-0.108465	1.000000	-0.200000	-0.108465	-0.108465	-0.108465	
Viruses_FAV	-0.108465	-0.108465	-0.200000	1.000000	-0.108465	-0.108465	-0.108465	
Viruses_HEV	-0.058824	-0.058824	-0.108465	-0.108465	1.000000	-0.058824	-0.058824	
Viruses_Hyopneumoniae	-0.058824	-0.058824	-0.108465	-0.108465	-0.058824	1.000000	-0.058824	
Viruses_Mavium	-0.058824	-0.058824	-0.108465	-0.108465	-0.058824	-0.058824	1.000000	
Viruses_PRRS	-0.058824	-0.058824	-0.108465	-0.108465	-0.058824	-0.058824	-0.058824	
Viruses_PRRSV	-0.058824	-0.058824	-0.108465	-0.108465	-0.058824	-0.058824	-0.058824	
Viruses_RNA	-0.058824	-0.058824	-0.108465	-0.108465	-0.058824	-0.058824	-0.058824	
Viruses_Rotavirus	-0.058824	-0.058824	-0.108465	-0.108465	-0.058824	-0.058824	-0.058824	
Viruses_SwinePox	-0.058824	-0.058824	-0.108465	-0.108465	-0.058824	-0.058824	-0.058824	
Viruses_TGE	-0.058824	-0.058824	-0.108465	-0.108465	-0.058824	-0.058824	-0.058824	
Viruses_Typhimurium	-0.058824	-0.058824	-0.108465	-0.108465	-0.058824	-0.058824	-0.058824	
Signs_Anorexia	-0.150414	-0.150414	-0.277350	-0.277350	0.391077	0.391077	-0.150414	
Signs_HPS	-0.058824	-0.058824	-0.108465	0.542326	-0.058824	-0.058824	-0.058824	
Signs_Haemorrhages	-0.085749	0.685994	-0.158114	-0.158114	-0.085749	-0.085749	-0.085749	
Signs_Haemorrhagic	-0.085749	-0.085749	-0.158114	-0.158114	-0.085749	-0.085749	0.685994	
Signs_LessFever	-0.129641	-0.129641	0.478091	0.478091	-0.129641	-0.129641	-0.129641	
Signs_Lethargy	0.685994	-0.085749	-0.158114	-0.158114	-0.085749	-0.085749	-0.085749	
Signs_Perirenal	-0.058824	-0.058824	0.542326	-0.108465	-0.058824	-0.058824	-0.058824	
Signs_RedSpots	-0.058824	-0.058824	-0.108465	-0.108465	-0.058824	-0.058824	-0.058824	

22 rows × 22 columns

Out[27]:

	opneumoniae	Viruses_Mavium	Viruses_PRRS	Viruses_PRRSV	Viruses_RNA	...	Viruses_TGE	Viruses_Typhimurium	Signs_Anorexia	Signs_HPS	Signs_Haemorrhag
	-0.058824	-0.058824	-0.058824	-0.058824	-0.058824	...	-0.058824	-0.058824	-0.150414	-0.058824	-0.0857
	-0.058824	-0.058824	-0.058824	-0.058824	-0.058824	...	-0.058824	-0.058824	-0.150414	-0.058824	0.6859
	-0.108465	-0.108465	-0.108465	-0.108465	-0.108465	...	-0.108465	-0.108465	-0.277350	-0.108465	-0.1581
	-0.108465	-0.108465	-0.108465	-0.108465	-0.108465	...	-0.108465	-0.108465	-0.277350	0.542326	-0.1581
	-0.058824	-0.058824	-0.058824	-0.058824	-0.058824	...	-0.058824	-0.058824	0.391077	-0.058824	-0.0857
	1.000000	-0.058824	-0.058824	-0.058824	-0.058824	...	-0.058824	-0.058824	0.391077	-0.058824	-0.0857
	-0.058824	1.000000	-0.058824	-0.058824	-0.058824	...	-0.058824	-0.058824	-0.150414	-0.058824	-0.0857
	-0.058824	-0.058824	1.000000	-0.058824	-0.058824	...	-0.058824	-0.058824	-0.150414	-0.058824	0.6859
	-0.058824	-0.058824	-0.058824	1.000000	-0.058824	...	-0.058824	-0.058824	0.391077	-0.058824	-0.0857
	-0.058824	-0.058824	-0.058824	-0.058824	1.000000	...	-0.058824	-0.058824	-0.150414	-0.058824	-0.0857
	-0.058824	-0.058824	-0.058824	-0.058824	-0.058824	...	-0.058824	-0.058824	0.391077	-0.058824	-0.0857
	-0.058824	-0.058824	-0.058824	-0.058824	-0.058824	...	-0.058824	-0.058824	-0.150414	-0.058824	-0.0857
	-0.058824	-0.058824	-0.058824	-0.058824	-0.058824	...	1.000000	-0.058824	0.391077	-0.058824	-0.0857
	-0.058824	-0.058824	-0.058824	-0.058824	-0.058824	...	-0.058824	1.000000	-0.150414	-0.058824	-0.0857
	0.391077	-0.150414	-0.150414	0.391077	-0.150414	...	0.391077	-0.150414	1.000000	-0.150414	-0.2192
	-0.058824	-0.058824	-0.058824	-0.058824	-0.058824	...	-0.058824	-0.058824	-0.150414	1.000000	-0.0857
	-0.085749	-0.085749	0.685994	-0.085749	-0.085749	...	-0.085749	-0.085749	-0.219265	-0.085749	1.0000
	-0.085749	0.685994	-0.085749	-0.085749	0.685994	...	-0.085749	-0.085749	-0.219265	-0.085749	-0.1250
	-0.129641	-0.129641	-0.129641	-0.129641	-0.129641	...	-0.129641	-0.129641	-0.331497	-0.129641	-0.1889
	-0.085749	-0.085749	-0.085749	-0.085749	-0.085749	...	-0.085749	0.685994	-0.219265	-0.085749	-0.1250
	-0.058824	-0.058824	-0.058824	-0.058824	-0.058824	...	-0.058824	-0.058824	-0.150414	-0.058824	-0.0857
	-0.058824	-0.058824	-0.058824	-0.058824	-0.058824	...	-0.058824	-0.058824	-0.150414	-0.058824	-0.0857

FIGURE 4: Correlation Matrix (Numerical Attributes)

B. BUILDING A HEATMAP FROM THE CORRELATION MATRIX

The next figure (Fig.5) shows a heat map created from the correlation matrix. The value 1 means that the correlation coefficient is very high, and the value ±0.05 means that the variables are in some way dependent on each other. Nevertheless, each character corresponds to a certain virus from the data set and vice versa. This correlation matrix shows the relationships between viruses and characters. For example, the signs lethargy, haemorrhagia and haemorrhages 0.69 are associated with the viruses 0.69 (Erysipela, Choleraesuis, Mavium, PRRS, RNA and Typhimurium) with a very high correlation coefficient of 1. The signs perirenal and HPS 0.54 are associated with the viruses 0.54 (FAV-4, Escherichia) with a very high correlation coefficient of 1. The sign less fever 0.48 is correlated with the viruses 0.48 (FAV-4 and Escherichia) with a very high correlation coefficient of 1. The sign Anorexia 0.39 is correlated with the viruses (TGE, Rotavirus, PRRSV, Hyopneumoniae and HEV) with a very high correlation coefficient of 1. The sign red spots -0.059 is correlated with the virus 0.059 (Swine Pox) with a very high correlation coefficient of 1.

Out[18]: <Axes: >

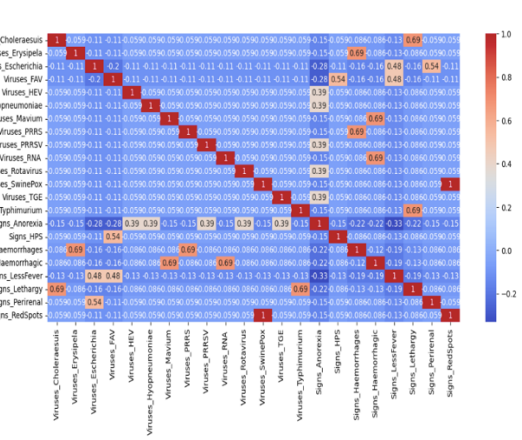

FIGURE 5: Heatmap from the Correlation Matrix

C. EVALUATION OF LINEAR REGRESSION MODEL PERFORMANCE

TABLE II summarises the model performance results obtained from linear regression.

TABLE II: SUMMARY OF THE EVALUATION OF THE LRM

R^2	MAE	MSE	Actual values	Predicted Values	Differences
1.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0

The R^2 value is comparable to accuracy. It gives a quick indication of how well the model works. In general, the closer the R^2 value is to 1.0, the better the model. In this study, the R^2 value gives us an indication of how accurate the prediction of this model is by establishing relationships between viruses and signs.

The MAE gives a better indication of how far our model's predictions deviate from the average. The values obtained were 0.0s. This underlines that there is a correlation between the predicted values and the actual values. The prediction using LR was 0.0 with 100 per cent accuracy.

Figure 6 shows the relationships between ASF viruses and clinical symptoms using the linear regression model with the metrics R-squared, MAE and MES.

Figure 7 shows a better result of the regression metrics used such as R^2, MAE and MSE. The R^2 value is 1.0, which proves that our model is accurate. The residual plots show us exactly how perfect our model is in terms of prediction, which is 0.0 with 100 per cent accuracy. MAE indicates that the predictions of the model are 0.0 on average. MSE is calculated by squaring the differences between the predicted values (PV = "Signs") and the actual values (AV =" Viruses"). MSE indicates 0.0.

D. CROSS VALIDATION AND SCORING PARAMETERS WITH REGRESSION METRICS

TABLE III: CROSS VALIDATION METRICS WITH REGRESSION MODEL

Cross_Validation_Accuracy	Cr_Val_Precision	Cr_Recall	Cr_F1_Precision
1.0	0.0	0.0	0.0
Cr_R^2	**Cr_MAE**	**Cr_MSE**	**Improvements**
1.0	0.0	0.0	The more data, the better.

The table (Table III) shows that the cross-validated precision is 1.0 with an accuracy of 100 per cent. The cross-validated precision is 0.0 with an accuracy of 100 per cent. The cross-validated recall is 0.0 with an accuracy of 100 per cent. The combined precision (prediction) was

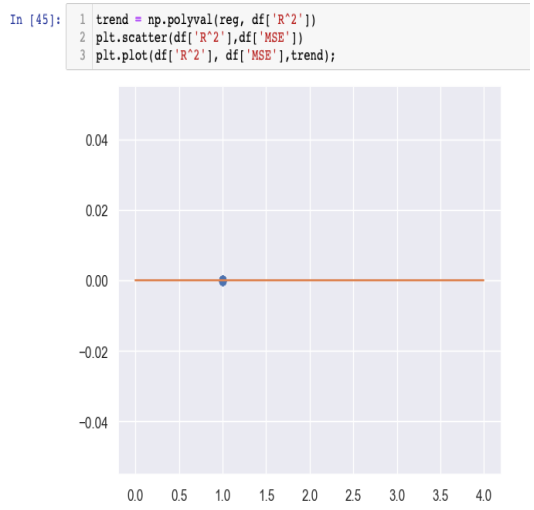

FIGURE 6: Evaluation metric with LR

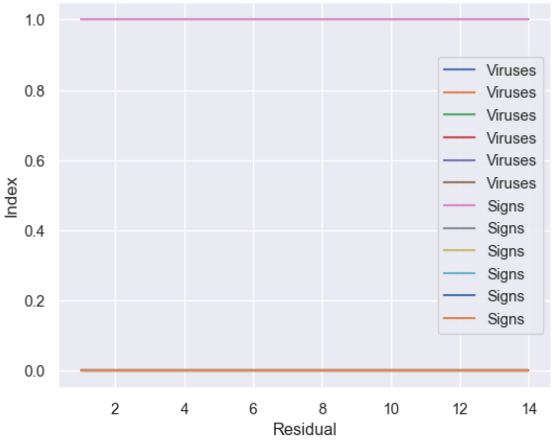

FIGURE 7: Residual

The residual plots show the linear relationship between the independent and the dependent variables.

Figure 7 shows that the three conditions for linear regression metrics are met. These metrics establish relationships between ASF viruses and its clinical signs.

0.0 with an accuracy of 100 per cent. Cr_R^2 was 1.0 with an accuracy of 100 per cent. Cr_ MAE and Cr_MSE were both 0.0 with an accuracy of 100 per cent. Therefore, the performance of the regression model is perfect for establishing relationships between viruses and signs.

E. EVALUATING THE PERFORMANCE OF RANDOM FOREST CLASSIFIER MODEL WITH CROSS-VALIDATION AND SCORING PARAMETERS(CLASSIFICATION METRICS)

TABLE IV: CROSS VALIDATION METRICS WITH CLASSIFICATION MODEL

Cross_Validation_Accuracy	**Cr_Val_Precision**	**Cr_Recall**	**Cr_F1_Precision**
1.0	0.0	0.0	0.0
Cr_R^2	**Cr_MAE**	**Cr_MSE**	**Improvements**
1.0	-0.0	-0.0	All score objects must be higher than lower values or negative.

The table (Table IV) shows that the cross-validated precision is 1.0 with an accuracy of 100 per cent. The cross-validated precision is 0.0 with a precision of 100 per cent. The cross-validated recall was 0.0 with a precision of 100 per cent. The combined precision (prediction) was 0.0 with a precision of 100 per cent. Cr_R^2 was 1.0 with a precision of 100 per cent.

However, the classification metrics with Cr_ MAE and Cr_MSE were -0.0, which led to a negative result. All evaluation objects follow the convention that higher return values are better than lower return values. Therefore, metrics that measure the distance between the model and the data, such as the metric mean_squared_error, are available as neg_mean_squared_error, which returns the negated value of the metric. In this case, the performance of the classification model is imperfect due to these negative values.

F. EVALUATING THE PERFORMANCE OF THE BAYESIAN REGRESSION MODEL

TABLE V: CROSS VALIDATION METRICS WITH BAYESIAN REGRESSION

Cross_Validation_Accuracy	**Cr_Val_Precision**	**Cr_Recall**	**Cr_F1_Precision**
1.0	0.0	0.0	0.0
Cr_R^2	**Cr_MAE**	**Cr_MSE**	**Improvements**
1.0	0.0	0.0	The more data, the better. Improve the current data

H. EVALUATING THE PERFORMANCE OF ANN MODEL

In cross validating the artificial neural network (ANN), we evaluated its performance using various metrics such as R2 value, MAE and MSE. The cross-validated accuracy was 1.0 with a precision of 100 per cent. The cross-validated accuracy was 0.0 with a precision of 100 per cent. The cross-validated recall was 0.0 with a precision of 100 per cent. The combination of precision (prediction) was 0.0 with an accuracy of 100 per cent. Cr_R^2 was 1.0 with an accuracy of 100 per cent. Cr_ MAE and Cr_MSE were both 0.0 with 100 per cent. Bayesian regression was tested with all the basic features. Interestingly, the metrics r2_score, MAE and MSE of the parameters are almost identical to those of the linear regression we discussed earlier. The cross-validated precision was 1.0 with a precision of 100 per cent. The cross-validated precision was 0.0 with a precision of 100 per cent. The cross-validated recall was 0.0 with a precision of 100 per cent. The combination of precision (prediction) was 0.0 with a precision of 100 per cent. Cr_R^2 was 1.0 with a precision of 100 per cent. Cr_ MAE and Cr_MSE were both 0.0 with an accuracy of 100 per cent. It can be concluded that the performance of the Bayesian regression model is perfect for establishing relationships between ASF viruses and clinical signs.

G. EVALUATING THE PERFORMANCE OF DECISION TREE MODEL

The decision tree regression model was tested on the ASF dataset to establish relationships between viruses and signs. The results are consistent with those of the linear regression and Bayesian regression models. The cross-validated accuracy was 1.0 with a precision of 100 per cent. The cross-validated accuracy was 0.0 with an accuracy of 100 per cent. The cross-validated recall was 0.0 with a precision of 100 per cent. The combined accuracy was 0.0 with a precision of 100 per cent. The Cr_R^2 was 1.0 with a precision of 100 per cent. Cr_ MAE and Cr_MSE were 0.0 with a precision of 100 per cent. We conclude that the decision tree regressor model is perfectly suited to establish relationships between ASF viruses and traits.

TABLE VI: CROSS VALIDATION METRICS WITH DECISION TREE

Cross_Validation_Accuracy	Cr_Val_Precision	Cr_Recall	Cr_F1_Precision
1.0	0.0	0.0	0.0
Cr_R^2	Cr_MAE	Cr_MSE	Improvements
1.0	0.0	0.0	The more data, the better. Improve the current data

accuracy. This proves that ANN is perfect for establishing correlations between ASF viruses and Signs.

I. PERFORMANCE OF THE MACHINE LEARNING MODELS

These models were trained with the ASF dataset. Three general metrics were used to evaluate the performance of the models, namely R2, MAE and MSE. These metrics provide different insights into the accuracy and predictive power of the model. For example, the performance metrics of the classification model using MAE and MSE were -0.0, which is a negative result. This means that the

test performance of the classification model was not sufficient to correlate ASF viruses with signs. Thus, the performance of this model was poor and unsuitable for the diagnosis of ASF. The MAE measures the average absolute difference between predicted and actual values. It provides a direct measure of the average magnitude of model error. Therefore, 0 means a perfect fit when the predicted values match the actual values perfectly. MSE measures the average squared difference between the predicted values and the actual values. As MAE, where 0 means a perfect fit if the predicted values match the actual values perfectly.

Linear regression, Bayesian regressor, decision tree regressor and ANN are the perfect models for ASF diagnosis. These models can establish correlations between ASF viruses and signs with accuracy. For example, the R^2 value ranges from 0 to 1, where 1 indicates a perfect match and 0 means that the model cannot establish correlations between ASF viruses and signs. In this study, the results show that the R^2 value and the R^2 value of cross-validation are 1.0 at 100 per cent. accuracy. The MAE and the MSE value were 0.0 at 100 per cent accuracy. The cross–validation precision was 0.0 at 100 per cent accuracy. The cross-validated recall value was 0.0 at 100 per cent accuracy. The combined precision (prediction) was 0.0 at 100 per cent accuracy. This proves that these models are perfect for ASF diagnosis. In this paper, we have discussed the multiple linear regression models and ANN which have shown better performance than other ML algorithms for the ASF dataset.

J. EVALUATION OF THE CAUSAL ML MODELS WITH REGARD TO THEIR PERFORMANCE (ACCURACY, PRECISION)

This section discussed the performance of the algorithms that gave better results. Significantly, the multiple polynomial regressions performed relatively better on the ASF data collected. The next table (Table VII) compares the performance of the ANN, linear regression (LR), Random Forest Classifier, Bayesian regression (BR) and decision tree (DT) models for the ASF diagnosis discussed.

TABLE VII: COMPARISON AND DIFFERENCES OF THE MODELS PERFORMANCE

MODELS	ACCURACY	PRECISION	DIFFERENCE	NOTES
LR	100%	100%	NO	Perfect Model for ASF diagnosis.
ANN	100%	100%	NO	Perfect Model for ASF diagnosis.
BR	100%	100%	NO	Perfect Model for ASF diagnosis.
DT	100%	100%	NO	Perfect Model for ASF diagnosis.
CLASSIFICATION	-0.0	-0.0	YES Negative scores	Imperfect Model for ASF diagnosis.

V. CONCLUSION

This study was initiated with a systematic literature review to solve a research problem. The concept idea was tested by developing a causal machine learning model capable of establishing small-scale relationships between ASF viruses and disease signs using historical laboratory data. The main goal of this research was to apply a causal ML model that extracts actionable information from ASF observation datasets to make intervention decisions for accurate ASF diagnosis.

An experimental method based on causal theory was proposed. The application of causal theory to experiments was evident in the development of an approach to evaluate the performance of the models using three commonly used metrics, namely R^2, MAE and MSE. These metrics give different insights into the accuracy and predictive power of the model. The application of causal theory in experimental research has highlighted the different dimensions of the applied research approach. Structural causal models (SCMs) have been used as a quantitative disease analysis to test the accuracy of ASF diagnosis.

The pragmatic validity of a generic design refers to whether it works after contextualization and implementation or not [23]. This research could be replicated in different contexts (types of diseases) to inductively generalise the findings. Different possible trait variables and ML algorithms were tested to build a better CML model. It was found that both multiple linear regression models (B, LR) and non-linear regression models (ANN, DT) performed very well and had a high level of accuracy and precision, often 100 per cent.

The results showed that the R^2 value and the cross-validation R^2 value = 1.0 at 100 per cent accuracy. The MAE value and the MSE value were both = 0.0 at 100 per cent accuracy. The cross–validated precision was = 0.0 at 100 per cent accuracy. The cross-validated recall was = 0.0 at 100 per cent accuracy. The combination of precision (prediction) was = 0.0 with 100 per cent accuracy. This proves that these models are perfectly suited for ASF diagnosis.

However, we also tried to evaluate the performance of the classification model for ASF diagnosis. Unfortunately, the results showed that the performance metrics of the classification model (Random Forest Classifier) gave a negative result with MAE and MSE = -0.0. This result shows that the test performance of the classification model is not sufficient to correlate ASF viruses with signs. Thus, the performance of this model ‘was poor and unsuitable for ASF diagnosis.

The size of the training dataset is crucial in this research as it is not possible to obtain a large ASF dataset in real-world scenarios. In this study, we developed a causal machine learning model that has acceptable inference accuracy and precision with a small ASF dataset to test our idea. The comparison of the CML model of the proposed approach with the existing systems shows that

the proposed CML model is excellent with 100 per cent accuracy and precision in establishing relationships between viruses and signs for ASF diagnosis.

The ideal training size of the ASF dataset for better model performance still needs to be explored. Further empirical studies are needed that consider laboratory tests, clinical signs and symptoms, field diagnoses, disease stages, expertise, and training with CML for ASF diagnosis. In addition, one of the future research directions is the following: Building a comprehensive inference model based on multiple ASF datasets.

REFERENCES

[1] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 8, pp. 1798–1828, 2013.

[2] A. Coates, A. Y. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in international conference on artificial intelligence and statistics, 2011, pp. 215–223.

[3] M. Anderson, D. Antenucci, V. Bittorf, M. Burgess, M. J. Cafarella, A. Kumar, F. Niu, Y. Park, C. Ré, and C. Zhang, "Brainwash: A Data System for Feature Engineering.," in Proc. CIDR 2013, 2013.

[4] P. Domingos, "A few useful things to know about machine learning," Commun. ACM, vol. 55, no. 10, pp. 78–87, 2012.

[5] I. Guyon and André Elisseeff, 2006. "An Introduction to Feature Extraction," in Guyon, Isabelle, Steve Gunn, Masoud Nikravesh, and Lofti A. Zadeh, eds. Feature Extraction: Foundations and Applications, pp. 1-25. Springer Berlin Heidelberg, 2006.

[6] H.-F. Yu, H.-Y. Lo, H.-P. Hsieh, J.-K. Lou, T. G. McKenzie, J.-W. Chou, P.-H. Chung, C.-H. Ho, C.-F. Chang, Y.-H. Wei et al., "Feature engineering and classifier ensemble for kdd cup 2010," KDD Cup, 2010.

[7] L. Breiman and J. H. Friedman, "Estimating optimal transformations for multiple regression and correlation,"

Journal of the American statistical Association, vol. 80, no. 391, pp. 580–598, 1985.

[8] S. Scott and S. Matwin, "Feature engineering for text classification," in ICML, vol. 99. Citeseer, 1999, pp. 379–388.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg et al., "Scikit-learn: Machine learning in python," The Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[10] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, ´ C. Olah, M. Schuster, J. Shlens, B. Steiner, I.

Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, ´ P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[11] What is the intuitive explanation of feature engineering in ML? - Quora, www.quora.com. Retrieved, 2023-07-07.

[12] S. Bird, E. Klein, and E. Loper, Natural Language Processing with Python. O'Reilly Media, Inc., 2009.

[13] H. Raghavan, O. Madani, and R. Jones, "InterActive Feature Selection," in Proc. IJCAI 2005, 2005, vol. 5

[14] C.S., Kampouridis, M., Freitas, A.A., Feature Engineering for Improving Financial Derivatives-based Rainfall Prediction, IEEE World Congress on Computational Intelligence, Vancouver, Canada (2016).

[15] F. Adafre and M. de Rijke. 2005. Feature engineering and post-processing for temporal expression recognition using conditional random fields. In Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing.

[16] Jeyanthi Narasimhan, Lawrence Holder, Feature Engineering for Supervised Link Prediction on Dynamic Social Networks, School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA 99164-2752, USA, 7 Oct. 2014.

[17] Kalyan Veeramachaneni, Una-May O'Reilly, Colin Taylor Towards Feature Engineering at Scale for data from Massive open Online Courses, arXiv:1407.5238v1 [cs.CY] 20 Jul 2014.

[18] Christopher Re, Amir Abbas Sadhgein, Zifei Shan, Jaeho Shin, Feiran Wang, Sen Wu, Ce Zhang, Feature Engineering for knowledge Base construction (KBC), Copyright 2014 IEEE.

[19] European Union Reference Laboratory for African Swine Fever (ASF)," asf-referencelab.info," available at

https://asf-referencelab.info/asf/en/procedures-diagnosis/diagnostic-procedures. [Accessed on 10/ 02/ 2023].

[20] R.R. Hocking "A Biometrics Invited Paper. The Analysis and Selection of Variables in Linear Regression". Biometrics. International Biometric Society, vol.32, no.1, pp.1.1976.

[21] A. Kassambara "Stepwise Regression Essentials in R - Articles - STHDA, STHDA". Available at: http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/154- stepwise-regression-essentials-in-r/ [Accessed: 16 May 2023]. 2018.

[22] L. Hansheng and V. Govindaraju "Regression time warping for similarity measure of sequence". The Fourth

International Conference on Computer and Information Technology, 2004. CIT '04. IEEE, pp. 826–830. doi:

10.1109/cit.2004.1357297.2004.

[23] J. Aken, A. Chandrasekaran and J. Halman "Conducting and publishing design science research: Inaugural essay of the design science department of the Journal of Operations Management", Journal of Operations Management. Elsevier B.V., vol. 47, no. 48.