ENHANCING THE DETECTION OF DEBIT CARD FRAUD DETECTION USING LOGISTIC REGRESSION AND RANDOM FOREST TECHNIQUES

***Abstract*--**Debit card fraud is one of the major financial crimes globally, causing a very great financial losses for financial institutions and individuals. The traditional mode of fraud detection systems often struggles to keep with the latest change in fraud patterns, due to the dynamism of the criminals resulting in high rates of false positives. This project proposes an improved system based on machine learning models to accurately and effectively identify fraudulent transactions.With machine learning models, fraudulent activities can be monitored and identified in real time. It is able to adapt to the changing nature or approach of fraudsters due to advancement in technology unlike the traditional model that is static in nature. Machine learning approach to fraud detection will mitigate the instances of false positive.This project focuses on utilizing machine learning algorithms, namely Random Forest (RF) and Logistic Regression for detecting debit card fraud. A series of rigorous experiments were conducted to evaluate the effectiveness of RF and LR in detecting debit card fraud. Evaluation is carried out using various performance metrics, including accuracy, precision, recall, sensitivity, specificity and F1-Measure.

***Keywords*--**debit card fraud, machine learning, random forest, logistic regression, fraud detection, class imbalance

1. Introduction

Fraud is a criminal action carried out to harm people. It involves the act of impersonation or manipulation of facts to deceive people in order to wrongfully acquire what is theirs. When it comes to finance, fraud is a deceptive way of gaining access to funds and assets owned by someone else. In addition to the huge losses experienced by financial institutions, public trust is also greatly affected by this criminal act. Traditional rule-based systems look for suspected fraudulent actions using predetermined rules and patterns. But these systems struggle to adjust to new and developing fraud strategies, which results in many false negatives and potential financial losses. [1]. The use of machine learning algorithms has drawn a lot of interest as a solution to these restrictions. One of the major things affecting the world's economy is fraud, the effects of it are spread across individuals, organizations, and even governments.

As more dependence on online transactions continues to grow because of the introduction and encouragement of cashless policy, fraudsters now find it even more convenient to carry out their evil acts. The emergence and increasing rise of online transactions has significantly impacted the financial sector making transactions to be more convenient and efficient. However, this technological advancement has also resulted in a surge of financial fraud incidents specifically related to the use of debit cards [2]. The transactions through the mode of debit cards are usually carried out online, and the fraud activities-imposed fear on the debit card owners [3]. Because the fraudsters can now from the comforts of their home gain access to people's card details without traces.

Debit card fraud is the action of a strange person carrying out a financial transaction by making use of someone else's card information without the permission of the owner and with no desire to return the investment made. Usually, debit card frauds occur when someone's card is stolen, lost, and sometimes the fraudsters collaborate with the Point-of-Sale Operators (POS) or bank staff. With cashless policy being enacted by the Federal Government of Nigeria and subsequently becomes a fast-growing transactional process, it becomes difficult to do away with by citizens. Due to rise in debit card fraud activities in Nigeria, many transactions involving debit/credit are being declined. Detecting this act is very important. Machine Learning approach provides a very useful way of tackling this problem. These techniques analyse a customer's data and detect patterns that may indicate fraudulent activities [4].

The sad news is that despite all the measures provided to put an end to fraudulent activities involving debit cards, fraudsters still have their way. Among the strategies fraudsters employed to commit fraud is that they usually make their behavioural activities to be unique and look like the real ones. Solutions to fraud can be classified into two, the prevention and detection methods. Fraud prevention is more of a proactive measure, it uses rules to stop fraud from happening. Technologies like Address Verification System (AVS) and Card Verification System (CVS) are usually operated to prevent fraud [5]. Fraud detection is the process of getting to know if a transaction is legit or not. It is therefore imperative to find techniques that is capable to detect these frauds despite their dynamism. This study focuses on the application of machine learning techniques in combating debit card frauds.

2. Literature Review

2.1 Review of Related Works

Due to the huge effects and surge of debit card fraud, researchers are constantly motivated to find a lasting solution to this menace. Quite a few methods have been introduced and tested. Some of which are reviewed below.

A study on preventing debit card fraud by [6] employed data from three distinct proportions of Debit card datasets. Because these datasets were severely skewed, an oversampling technique was used to balance them, after the usage of the resampling approach. The performance of three machine learning algorithms Logistic Regression, Naive Bayes, and KNN is tracked with a comparison of their results. The performance of the algorithms is assessed in terms of accuracy, sensitivity, specificity, precision, F-measure, and area under the curve. These results demonstrated that the logistic regression fraud prediction model outperforms models built around Naive Bayes and K-nearest Neighbour. Using under-sampling techniques on the data prior to creating the prediction model also produces better outcomes.

[7]created a model that is tested on the German Debit Card Dataset utilizing the J48, Decision Tree, Adaboost, Random Forest, Naive Bayes, and PART algorithms. The Filter and Wrapper Approaches have increased the accuracy of J48 and PART. Random Forest obtained an accuracy of 76.4% with and without the feature selection approach which was the greatest. With the wrapper approach, Naive Bayes obtains an accuracy of 75%.

[8]applied supervised techniques and algorithms are utilized to detect fraud and produce findings that are roughly correct. In this study, an isolation forest algorithm is used to classify the data sets in order to look for fraud activities. The data sets were obtained from reputable survey companies. For evaluation, the binary classification method will be preferred. Whenever a transaction has taken place, and someone has attempted to engage in any fraudulent activity. By using classification techniques, the system's performance was compared. The isolation forest approach is utilized for outliers, and for all categorized datasets, the accuracy obtained is 99.87%.

[9]. The methods were applied to the raw and previously treated data. The findings of the investigations indicate that Random Forest, with an exact precision of 98.6%, has the best outcomes, followed by Logistic Regression, SVM, and Decision Tree with exact precisions of 97.5%, 97.7%, and 95.5%, respectively. Using the dataset provided by ULB, the results show that Random Forest has the highest precision and accuracy of 98.6% in the challenge of detecting debit card fraud.

[10]used two different types of random forest algorithms to prepare the characteristics of common and unusual transactions. The two arbitrary random forest methods that are used are taken into consideration, and their presentations on detecting Debit card fraud are investigated. The setup for the two random-tree-based random forests and the CART-based random forests methods, comes from bootstrapped experiments. Using various datasets with various dataset proportions, the three experiments were run for the two methods. The performance of these algorithms was assessed throughout all three tests using additional measures, including the change in the rate of transactions and the mean rate of the model. In each test, the cart-based random forest outperformed the others.

In the work of [11], they used a real-world data set, they applied supervised machine learning techniques, leveraged those algorithms to create a super classifier using ensemble learning, and then compared the effectiveness of the supervised methods with their super classifier implementation. Ten machine learning techniques, including Random Forest, Stacking Classifier, XGB Classifier, Gradient Boosting, Logistic Regression, MLP Classifier, SVM, Decision Tree, KNN, and Naive Bayes, were used to compare the output of their super classifier to the accuracy, recall precision, and confusion matrix. As a result, they discovered that Logistic Regression is more effective in predicting fraud transactions.

[12]compared the results of two random forests. CART-based random forest, random forest based on random trees. To train the behaviour aspects of regular and abnormal transactions, they employ various random forest algorithms, each of which has a different performance and base classification. Using the dataset of a Chinese e-commerce company, they ran both algorithms. The subsets' fraud transaction ratio ranges from 1:1 to 10:1. The accuracy of the random-tree-based random forest is therefore 91.96%, whereas the accuracy of the CART-based random forest is 96.7%. Several issues, such as uneven data, have arisen as a result of the data being from the B2C dataset. Thus, the algorithm can be enhanced.

[13]developed a method to predict phishing websites using Support Vector Machine (SVM) and multi-class classification based on association rule

techniques. Their approach aimed to enhance the accuracy of phishing detection systems. The study demonstrated that combining SVM with association rule techniques significantly improved the identification of phishing websites. This research contributes to cybersecurity by providing a robust model for detecting phishing threats.

[14]proposed numerous machine learning algorithms and evaluated them in relation to techniques for detecting debit card fraud. Among the several machine learning methods are Logistic Regression, Naive Bayes, Random Forest, and Multilayer Perceptron. An artificial neural network called a multilayer perceptron (ANN), which has four hidden layers and depends on relu activation to prevent negative values, is employed in this situation. Adam is employed as the performance optimizer. Because of this, the Logistic regression model's accuracy score is 97.46%, and the data set includes 56962 samples, 98 of which are fraudulent transactions. For the same dataset, Naive Bayes and Random Forest both achieved accuracy scores of 99.23% and 99.96%. As we can see, random forests produce the greatest results when it comes to debit card fraud detection. The ultimate accuracy for ANN was 99.93%.

[15]investigateda few classification techniques utilizing various metrics to evaluate various classifiers. Several ensemble and classification techniques were used to create the models. Ensemble models, support vector machines, decision trees, random forests, and logistic regression were among the models used. After these models underwent training, results were attained. The actual dataset produced equally excellent results, and compared to other classifiers, the Random Forest classifier outperformed them with a recall rate of approximately 96%.

3. Methodology

3.1 The Architecture of the Proposed Debit Card Fraud Detection Model

The conceptual view of the detection model is arranged into Three (3). The first block in the architecture is the Data collection and pre-processing phase, the second phase is the Model training and validation phase, and the last phase is the Model evaluation and result. Debit card data obtained is normalized in the first phase using the min-max normalization method to adjust the debit card fraud feature values on different scales to a comparable value. Thereafter, the dataset is sent to the feature selection technique.

Thereafter, the normalized data is passed to the second phase where the data is split into train and test sets based on a reasonable percentage. The trained models are evaluated using the train and test set, and the test result is displayed at the model evaluation and result phase.

3.2 Data Collection and Description

The dataset used in this work is a 143MBComma Separated Value (CSV) file extracted from Kaggle. The dataset was published in 2016. This dataset was based on two days' worth of transactions gathered from European cardholders in September 2013. The datasets have served as a very useful resource for researchers since its release in modelling different systems.

The dataset is exceptional when compared to other debit card fraudulent datasets used in the prior studies as it contained new features for identifying fraud activities. There are 284807 records in the dataset with all numeric data values.The dataset contains 31 features. It contains various information such as time, transaction amount, class, and other relevant information to debit card transactions encoded as V1 to V28 to safeguard sensitive information.

Both the train and test sets are in comma separated values (CSV) format where the debit card fraud instances are presented in rows and the features that describe the instances are represented in columns.

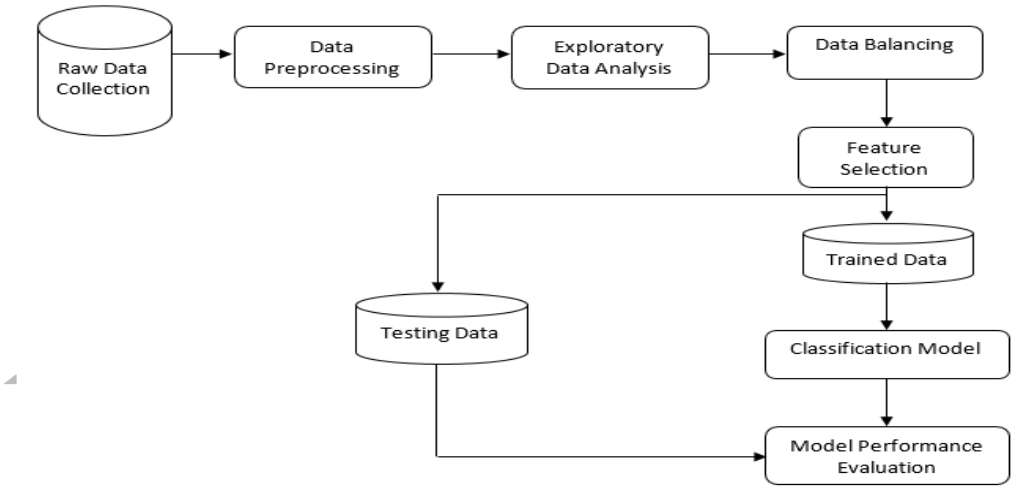

Figure 1: Proposed Model Architecture

3.3 Data Pre-processing

This phase guarantees the dataset's suitability for the intended experimental application by cleansing it of any irrelevant or redundant elements, filling in gaps, and adjusting unbalanced figures to align with the demands of machine learning algorithms. The variety of preprocessing techniques available is vast, yet their deployment is dependent upon the dataset's unique characteristics. Through Exploratory Data Analysis (EDA), deeper insight was gained into the data, which indicated a need for normalization to correct the scale of the features.

3.4 Data Normalization

Data normalization is applied to transform variant values to be between 0 and 1. Attribute value variation is often a challenge for most machine learning algorithms as they tend to favour attributes with high values over small attributes, particularly during feature selection processing. Hence, data normalization is employed in this research to handle attribute value variation among the feature values. This step ensures improved feature selection and model classification processes. Among the several normalization techniques available, the min-max normalization method was chosen to normalize the skewed debit card fraud dataset.

3.5 Selection of Debit Card Fraud Attack Features

In order to correctly select relevant attack features, feature selection was applied to the initial attack features. This procedure is to improve detection and reduce model computation complexity. It is worthy of note that the dataset used in this research contains 30 features and using all the features could result in resource complexities. Besides, not all the features might be relevant. However, to ensure relevant and informative features were selected, feature importance method was applied. The purpose of choosing this approach is to rank the feature base on the information each feature contained.

3.6 Data Balancing

The debit card fraud dataset obtained for this experiment was discovered to be imbalanced. In datasets where the class examples are not proportionately distributed, a significant disparity exists. A predominant class overshadows with a vast number of instances, while a subordinate class is scarcely represented, sometimes making up less than 1% of the dataset. Such disparities present difficulties in machine learning, especially within classification tasks, as standard algorithms are designed to expect uniform class representation. Consequently, this often results in high accuracy for the dominant class but inadequate prediction for the crucial, less represented class. To mitigate these issues, two resampling strategies was applied: Oversampling and Under-sampling. Resampling allows different classes to have about the same influence on the classification model's outputs by making the training data more balanced [16].

Under-sampling Method: A strategy used to handle class imbalance by reducing the number of instances in the majority class. One of the challenges confronting this approach is the tendency of losing relevant features that would lead difficulties in prediction during the reduction of the instances.

Oversampling Method: The approach of oversampling is the opposite of under-sampling. It solves data imbalance by increasing the minority class through replication or generation of new examples.

3.7 Random Forest

A random forest is a meta-estimator that fits several decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. A random forest algorithm was designed to curb the shortcomings found in standalone decision tree algorithms. A group of decision tree models called Random Forest (RF) are assembled to improve prediction accuracy and avoid over-fitting. Every tree is taught with a different collection of data.

3.8 Logistic Regression

Logistic regression is the second classifier employed for this research. It is mostly applied for prediction and classification problems. Logistic Regression is chosen because of its high accuracy on binary datasets. The objective of Logistic regression is to categorize the fraud and non-fraud classes in the dataset by using a linearly separable line.

3.9 Evaluation Metrics

The performance evaluation of the proposed model was analysed based on the most relevant and widely used metrics for debit card fraud detection. The following measures would be considered: Accuracy, precision, Recall/sensitivity, and F1-score. The debit card fraud detection models were evaluated using standard metrics such as accuracy, precision, recall/sensitivity, and F1-Score.

Accuracy: This calculates all the transactions that were made. It divides the total number of correct transactions by all the transactions. Correct decision comprises of true positives (TP) and true negatives (TN). While the entire predictions consist of the positives (P) and negative (N) i.e. true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The mathematical expression is:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

Sensitivity: It measures the likelihood of getting a positive result when a test is conducted. It is also known as True Positive Rate (TPR). For example, testing for fraudulent activities. A high sensitivity means indicating a possibility of the test detecting fraud. While a low sensitivity means the test is less likely to detect fraudulent acts if it is present. High sensitivity makes the model better.

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (2)$$

Specificity: Known as True Negative Rate (TNR). It measures the possibility of getting a negative outcome. A high specificity shows that the model identifies the negative results correctly. While a low specificity means the model tends to take negative for positive.

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (3)$$

Precision: Measures the percentage of optimistic (fraud) predictions that are accurate. It divide the number of true positive results by the number of predicted positive results.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (4)$$

F-Measure: It is employed to assess how well a model performs. It combines both precision and recall into a singular score. It can be understood to be a weighted average of the precision and the recall, with 1 being the best and 0 as the worst.

$$\text{F-Measure} = \frac{2(Precision \times Recall)}{Precision+Recal} \quad (5)$$

4. Results and Discussions

4.1 Experiment 1

In this model, the oversampling approach was used, and is applied to the logistic regression classifier and random forest classifier, result of each model is compared in the table below and the graph for visualization.

Table 1: Evaluation/Measurement of Logistic Regression (LR) and Random Forest (RF) Classifier Oversampled method.

Parameters	LR	RF
True Positive	55537	56867
True Negative	51347	56853
False Positive	1336	6
False Negative	5506	0
Accuracy	93.984	99.995
Recall	90.98	100
Precision	97.65	99.99
F-Measure	94.12	99.99
Specificity	97.46	99.99

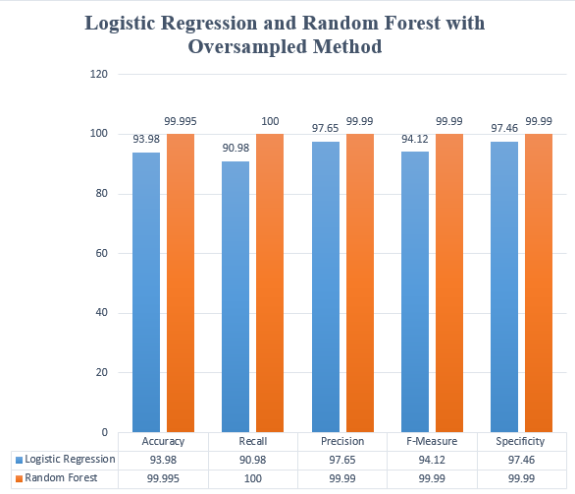

Figure 2: Logistic Regression and Random Forest with Oversampled Method

From Table 1, Random Forest classifier achieved a better accuracy, recall, Precision, f-measure, and specificity over Logistic Regression across the dataset.

4.2 Experiment 2

In this model, the under-sampling approach was used, and is applied to logistic regression and random forest classifier. The result of each model is compared in the table below and the graph for visualization as shown below

Table 2: Evaluation/Measurement of Logistic Regression (LR) and Random Forest (RF) Classifier under-sampled method.

Parameters	LR	RF
True Positive	98	95
True Negative	81	91
False Positive	0	3
False Negative	18	8
Accuracy	90.86	94.42
Recall	84.48	92.23
Precision	100	96.94
F-Measure	91.59	94.53
Specificity	100	96.81

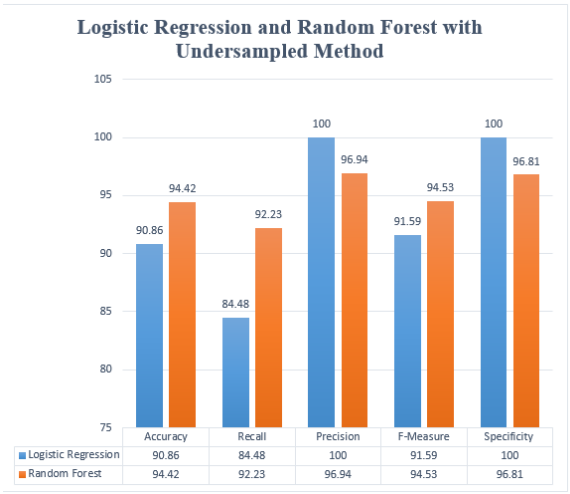

Figure 3: Logistic Regression and Random Forest with under-sampled Method

From Table 2, the random forest classifier performed marginally better than the logistic regression classifier, achieving a better accuracy of 94.42% compared to the logistic regression classifiers accuracy of 90.86%. Both demonstrated good recall, precision, f-measure, and specificity values, with the Random forest classifier exhibiting higher recall (92.23% against 84.48%) and f-measure (94.53% against 91.59%).

5. Conclusion

Debit card fraud continuously remain one of the growing challenges as perpetrators continue to come up with new ways of carrying out this criminal act. Because of the dynamism of the criminals, it is of high importance that matching systems are put in place to combat the menace. In building a predictive model for fraud detection, it is important to consider the changing nature of the fraudsters approach. However, this will require a lot of data. Because of advancement in technology, we find many and advanced ways of protecting assets so also fraudsters finds many advanced ways to carry out their attacks. As a result, we need to address the issue from the perspective of a protector and fraudster. When we understand how attacks are being done, it will become easier for us to detect and prevent it.

For this research, we examined two classifier techniques: Logistic Regression (LR) and Random Forests (RF). The focus of this research was on addressing the common occurrence of debit card fraud in financial institutions through binary classification. As shown in the result with the comparison of the investigation, Random Forest (RF) was the best technique for classifying debit card

fraud issues. It shows a better performance across various evaluation metrics for both under-sampled and oversampled dataset.

Overall, the techniques used in this work shows great potential in identifying genuine and fraudulent transactions.

5.1 Future Work

Based on the results and findings from this study, here are some few recommendations for future works.

i. Usage of more/robust datasets for more detailed research in the problem domain as more datasets determines the efficiency of the models.
ii. Examine the usage of ensemble models, like k-nearest-neighbour (KNN), Support Vector Machine (SVM), and eXtreme Gradient Boosting (XGBoost) to combine the advantages of many methods and raise the accuracy of fraud detection.
iii. Application of Deep leaning techniques such as Convolutional Neural Network and Artificial Neural Network

Disclaimer (Artificial intelligence)

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.) and text-to-image generators have been used during the writing or editing of this manuscript.

6. References

[1] P. Kumar, "Malicious Code Detection Using Machine Learning," International Journal for Research in Applied Science and Engineering Technology, vol. 12, no. 2, pp. 298-304, 2024.

[2] Bin Sulaiman, R., Schetinin, V., &Sant, P., "Review of Machine Learning Approach on Credit card Fraud Detection," Human-Centric Intelligent Systems, vol. 2, pp. 55 - 68, 2022.

[3] Arun, G.K., &Venkatachalapathy, D.K., "Convolutional Long Short Term Memory Model for Credit Card Detection," in 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020.

[4] Ileberi, E., Sun, Y., & Wang, Z. , "A Machine learning based credit card fraud detection using the GA algorithm for feature selection," Journal of Big Data, vol. 9, no. 1, 2022.

[5] R. Shakya, "Application of Machine Learning Techniques in Credit Card Fraud Detection," UNLY Theses, Dissertation, Professional Papers, and Capstones, 2018.

[6] Itoo, F., Meenakshi, & Singh, S. , "Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection.," International Journal of Information Technology, vol. 13, p. 1503–1511, 2020.

[7] Ajeet, S., & Jain, A., "Adaptive credit card fraud detection techniques based on feature selection method.," Advances in Computer Communication and Computational Sciences , pp. 167-178, 2020.

[8] Megasari, G., Jacky, C., Rianti, S., Phong, T., and Shankar, K. , "Machine Learning Methods for Analysis Fraud Credit Card Transaction," International Journal of Engineering and Advanced Technology (IJEAT), vol. 8, no. 6S, p. 2249 – 8958, 2019.

[9] Khare, A., Saxena, A., & Kumar, S., "The challenge of detecting debit card fraud," Journal of Financial Crime, vol. 25, no. 3, pp. 567-582, 2018.

[10] Xuan, S., Liu, G., Li, Z., Zheng, L., Wang, S. and Jiang, C., "Random Forest for Credit Card Fraud Detection," in IEEE 15th International Conference on Networking, Sensing and Control, Zhuhai,, 2018.

[11] Sahil, A., Kumar, R., & Sharma, P., "Using supervised machine learning techniques and ensemble learning to create a super classifier for fraud detection.," in In Proceedings of the 2020 International Conference on Machine Learning and Cybernetics (ICMLC), 2020.

[12] Shi, Y., Zhang, L., Wang, H., & Li, J., "Comparison of CART-based Random Forest and Random-Tree-based Random Forest for Fraud Detection," in In Proceedings of the 2019 International Conference on Data Science and Advanced Analytics (DSAA), 2019.

[13] Woods, N. C., Agada, V. E. and Ojo, A. K., "A Predicting Phishing Website Using Support Vector Machine and Multi-Class Classification Based on Association Rule Techniques," University of Ibadan Journal of Science and Logics in ICT Research, vol. 2, no. 1, pp. 28-39, 2018.

[14] Dejan, M., Ivan, S., & Marko, P., "Evaluation of Machine Learning Algorithms for Debit Card Fraud Detection," in Proceedings of the 2020 International Conference on Machine Learning and Cybernetics (ICMLC), 2020.

[15] Mishra, A., Singh, R., & Patel, S., "Evaluation of Classification Techniques for Fraud Detection.," in Proceedings of the 2021 International Conference on Data Science and Advanced Analytics (DSAA) , 2021.

[16] Bagui, S., & Li, K., "Resampling imbalanced data for network intrusion detection datasets," Journal of Big Data, vol. 8, no. 1, 2021.