Predicting Fraud in Mobile Money Transactions using Machine Learning: The Effects of Sampling Techniques on the Imbalanced Dataset

Francis E. Botchey[1,2], Zhen Qin[1], Kwesi Hughes-Lartey[1,2] and Kwame .E. Ampomah[1]
[1] School of Information and Software Engineering, University of Electronic Science and Technology of China
Chengdu 610051, China
E-mail: botcheyfrancis@gmail.com and https://en.uestc.edu.cn, qinzhen@uestc.edu.cn,
kwesihl@gmail.com, ampomahke@gmail.com
[2] Koforidua Technical University, Koforidua, Ghana
https://www.ktu.edu.gh

Keywords: ADASYN, fraud, mobile money, SMOTE

Received: May 30, 2020

Mobile Money Fraud is advancing in developing countries. We propose a solution to this problem based on machine learning. Labeled data from financial transactions which includes mobile money transactions are however, skewed towards the legitimate transactions. Machine learning models built with such skewed datasets are unreliable as the prediction algorithms will be biased towards the legitimate transactions. We investigate the performance of different sampling and weighting techniques such as Adaptive Synthetic Sampling (ADASYN) and Synthetic Minority Oversampling Technique (SMOTE). We select Logistic Regression for the experiments due to its simplicity and relatively low computational needs. The performance is evaluated with different metrics. Manually tuning the weights of the classes achieved the best results in our experiments.

Povzetek: Opisana je metoda za detekcijo prevar v mobilnih transakcijah s pomočjo strojnega učenja na neuravnoteženih podatkih.

1 Introduction

The use of mobile devices have become a rudimentary part of our daily lives. The way we conduct our daily activities have become heavily dependent on mobile devices. One significant aspect of our interactions with mobile devices that cannot be overemphasized is the way we conduct financial transactions. Financial technology, often referred to as Fintech is the use of innovations and technology that attempts to contend with the conventional way of undertaking financial transactions. Having the reach to conventional financial services, or being financially included, give opportunities and capabilities to individuals on how to plan, save, and stabilize their financial lives[1]. In most part of the developing world, access to formal financial services becomes virtually impossible as the infrastructure and services needed for formal financial inclusion are nonexistent. Where these financial infrastructure exist, often, customers have to travel long distances in order to access these services culminating in additional cost to the already impoverished individual. The implication of financial exclusion is that individuals with no access to conventional financial services tend to be poor and this is vividly evident in most developing countries.

Mobile Money Transactions (MMTs), are financial services offered by Telecommunication companies often refereed to as Mobile Network Operators (MNOs) that enable the transfers of funds (cash). These transfer of funds otherwise known as mobile money (MM)are offered between service subscribers (customers) and MNOs through the use of telecommunication channels[2]. From Demirgüç-Kunt et. al. [3], a third of all account holders which is 12% of the adult population reported having a mobile money account in sub-Saharan Africa. This comes as a relief since it provides financial inclusion for millions of people in developing economies.

MMTs, are fundamentally deployed using short message services (SMS) and Unstructured Supplementary Service Data (USSD) code which makes it very easy for the service to be deployed in rural areas with less accessibility to the internet. It also enable customers to use feature phones which are less expensive compared to their smart phone counterpart. However, Mobile Money Services can also also be deployed on smart phones using specialized mobile applications.

With its humble inception as M-Pesa in Kenya, MMTs have made huge in-roads into making people in developing countries financially inclusive. For example, in Ghana, Cote D'Ivoire, Benin, and Senegal, 54% of the combined adult population use MMTs on a regular basis[4].The value of MMTs is estimated to be $129.29 billion by 2021 across the globe according to Deloitte as cited by [5]. These tremendous gains, made by MMTs are on the verge of been eroded as fraudsters have been perpetuating fraud on the account of legitimate users. According to Busuulwa and Laryea cited by [5], in 2015, fraudulent transactions stood at 53% of the entire mobile money transactions in Uganda, 42% in Tanzania, 12% in Kenya and 23% in Ghana. This may be partly due to inadequate formal education, as the researchers observed as part of their studies, the willingness of MMT account holders to release their secret codes and other sensitive information to third parties with the aim of seeking help to undertake basic transactions.

Traditionally, there have been many approaches in dealing with fraud in financial transactions. These methods have been rule-based, data mining and other statistical methods. These methods, however, are gradually becoming unreliable as the known patterns and mode of operations of criminals gets sophisticated by the day.

The use of machine learning algorithms in predicting fraudulent transactions have witnessed an ascendancy. These algorithms, be it supervised or unsupervised, such as K-nearest neighbor, Naïve Bayes, logistic regression, and support vector machines(SVMs) are trained with data and are used after the training process to classify and predict financial transactions into legitimate and fraudulent ones. Other deep learning methods such as artificial neural networks (ANNs), and Convolutional Neural Networks (CNNs) have also been employed to detect anomalies in financial transactions. These Deep learning and Machine Learning Algorithms have shown high levels of accuracy in their predictions.

Given any dataset on financial transactions such as MMTs, the number of fraudulent transactions (positive class) compared to the legitimate (negative class) ones, constitutes a very small percentage of the dataset. This makes the datasets highly imbalanced [6] and predictions from such data using machine learning algorithms are skewed towards the legitimate transactions with the long term effect that predictions made with such data can be misleading. We select Logistic Regression as the machine learning algorithm for this work as it has proven its potency [7] in a multitude of fields for classification and prediction. It has been used in medicine[8, 9, 10], Engineering [11, 12], sports[13], Finance[14, 15, 16], computer science[17] etc. It is in this paradigm, that this paper explores the effects of different undersampling, weighting and oversampling techniques of equalizing the imbalanced dataset. These attempts to eliminate the problem of machine learning models whose results are lop-sided towards the majority class. Different undersampling and oversampling techniques are performed to evaluate the effects the imbalanced dataset have on predicting fraud in mobile money transactions. To the best of our knowledge none has been proposed. Our main contributions in this paper are in three folds:

- A proposed weighting technique to eliminate the bias effects imbalanced dataset have on machine learning algorithms.
- A fraud prediction model based on the proposed technique above to predict fraud in mobile money transactions as well as other financial transactions with imbalanced dataset.
- An in depth evaluation on the performance of our proposed model as well as that of the other analyzed models.

The remainder of this paper is structured as follows. In section II, we undertake a review of related works in the field of machine and deep learning. Section III gives a brief insight into machine learning and describe the foundations of our chosen machine learning algorithm; logistic regression. Section IV describes our dataset, our methods, and the experimental setup. In section V, we evaluate the performance of our models and discuss the results. We conclude the paper in section VI.

2 Related works

A survey of the majority of the studies done in the field of finance with regards to fraud prediction and detection using artificial intelligence, data mining and other statistical methods have focused on credit card fraud and others related to traditional banking activities.

An example is the work of[18]. In their narrative of financial fraud, the itemized list of financial fraud included only bank, corporate and financial fraud. Banking fraud decomposed further into credit card, money laundering, and mortgage fraud without a mention of MMTs, due to, perhaps its little prominence in the developed world.

In the remainder of this section, we briefly review related literature on data imbalance, supervised machine learning algorithms for classification and the evaluation metrics used.

Data imbalance. This situation arises when the dataset been used to train a machine learning algorithm for classification or other purposes is unevenly distributed between the positive and negative classes. According to[19], the percentage of fraud in audited financial report in of all the United States of America (positive class) was 0.6% compared to 99.4% which constitutes legitimate transactions (negative class). Models developed with such imbalance data often results in misclassification. To correct this anomaly, researchers either attempts to reduce the length of the negative class so it can be at par with that of the positive class. This is known as undersampling. Another method is to increase the length of the positive class with synthetic data using methods such as Synthetic Minority Oversampling Technique (SMOTE)[20] and Adaptive Synthetic Sampling (ADASYN)[21, 22] which are collectively known as oversampling.

Supervised classification Machine learning algorithms. We discuss recent works with classification algorithms such as Logistic Regression, Decision Trees, Naïve Bayes, K nearest neighbors (KNN) and other related algorithms.

Ref.[23] did a comparative analysis on the performance of Naïve Bayes, K-nearest neighbor, and logistic regression models in binary classification of imbalanced credit card fraud data. Their work analyzed the performance of these algorithms in classifying ULB dataset and proposed the use of other sampling techniques in relation to the imbalance data, having observed the fact that the nature of the dataset used had a serious impact on the obtained results.

The work of[24] looked at different learning algorithms with Université Libre de Bruxelles, Brussels, Belgium (ULB) dataset using SMOTE as the oversampling technique. They concluded by reporting on the performance of the classifiers based on the confusion matrix, recall, accuracy, and precision.

In the article[25], “Horse Race Analysis in Credit Card Fraud”, the researchers also considered Deep Learning, Logistic Regression, and Gradient Boosted Tree. They found out from their investigations by examining the Area Under the Curve(AUC) Receiver Operating Characteristics(ROC) values that, deep learning methods had the most powerful predictive power. The work, however, used undersampling which had the potential of discarding a large chunk of relevant information about the dataset. Several work has been done in the field of fraud detection using artificial neural networks[26, 27]. Neural networks, however, require large computational power as well as a huge dataset[28].

Evaluation metrics for machine learning algorithms. Different evaluation metrics may be used for evaluating the accuracy of different algorithms based on the uniqueness of their circumstances. The following are considered. **Classification Accuracy**. This is the ratio of the number of correct predictions to the total number of samples imputed for training and testing phase, **Confusion Matrix** which gives a vivid description of the performance of models, **AUCROC**[29, 30, 31, 32, 33] which is the probability that a machine learning algorithm will rank a randomly chosen positive example higher than a randomly chosen negative one, **F1-Score**[30], and **Root Mean Squared Error**. The F1-Score is normally used to predict since it has the ability to represent both the precision and recall[30].

An analysis of the reviewed literature creates the impression that majority of the data used by the researchers in developing their classifier models did little or nothing to address the problem of data imbalance. Again depending on the distinctiveness of environments, different evaluation metrics may be used in evaluating the performance and appropriateness of a model. For example, using just the model accuracy of a classifier as the performance criteria of an imbalanced dataset might give a wrong impression on its performance, the model accuracy might be very high but the true positive rate(TPR) and true negative rate(TNR) might be very low. The false positive rate (FPR) and false negative rate (FNR) may be very high which are indications of a bad classifier. These, therefore, leaves a sense of vagueness in the reported performance of these reviewed models as the imbalance problem was not properly handled and appropriate evaluations metrics used.

This paper investigates the performance of logistic regression in lieu of the reviewed work by experimenting with different undersampling, weighting and oversampling techniques to classify and predict fraud in Mobile Money Transactions.

3 Machine learning

Artificial Intelligence (AI) is a field under computer science which emphasizes on the creation of intelligent machines that work and behave like humans. Machine Learning is a sub field of AI where the concept has been defined differently by different school of thoughts. The classical definition by a pioneer in the field of AI, Arthur Samuel coined from his paper [34] is, “a field of study that gives computers the ability to learn without being explicitly programmed ". Tom Mitchell [35] also defined a well posed learning problem as: “ a computer program is said to learn from experience E with respect to some task T and some performance P, if its performance on T, as measured by P, improves with experience E ". Others have defined ML as the science of design and use of complex algorithms that has the ability to iterate over large datasets and analyze hidden patterns in the datasets. This process enables the machine to respond to different situations for which they have not been explicitly programmed to.

There are three categorizations of ML namely; supervised learning which uses labeled data for its training and testing having LR, Neural Networks and Support Vector Machines as examples, unsupervised learning which uses unlabeled data having self-organizing maps and and one-class support vector machines as examples and reinforcement learning which uses software agents to interact with the environment to learn from it while giving rewards and punishments.

3.1 Logistic regression (LR)

The mathematical foundation of LR is established on the sigmoid function shown in Fig. 1.

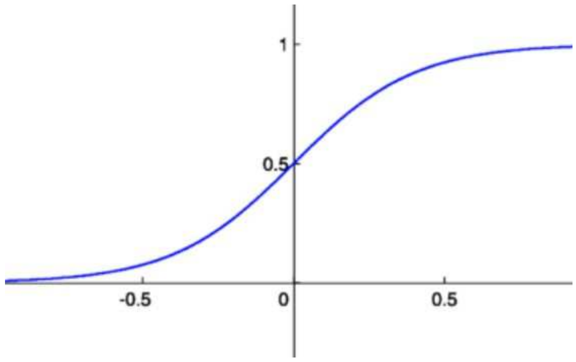

Figure 1: plot of a Sigmoid Function.

In LR we aim to achieve the output

$$0 \leq h_\theta(x) \leq 1 \tag{1}$$

where $h_\theta(x)$ is the hypothesis of LR

$$h_\theta(x) = \frac{1}{1 + \exp^{-\theta^T x}} \tag{2}$$

Given a training set, we fit the parameters θ. into equation (2), the probability output is given by equation(3).

$$h_\theta(x) = P(y = 1|x; \theta) \tag{3}$$

which implies that

$$P(y = 0|x; \theta) = 1 - P(y = 1|x; \theta) \tag{4}$$

Thus, if

$$h_\theta(x) \geq 0.5, predict 1 \tag{5}$$

and if

$$h_\theta(x) \leq 0.5, predict 0 \tag{6}$$

For a training set of M samples,$\{(x^1, y^1), (x^2, y^2), ...(x^m, y^m)\}$, it can be represented by a feature vector $x\epsilon \begin{vmatrix} x_0 \\ x_1 \\ \dots \\ x_n \end{vmatrix}$, Where the first parameter, $\theta_0 = 1$ and $y\epsilon\{0, 1\}$
For the hypothesis in equation(2), the cost function can be deduced as

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^i + (1 - y^i)\log((1 - h_\theta)(x^i))\right] \tag{7}$$

The parameter θ, in this paper was fitted by minimizing equation (7) using Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm, an algorithm for parameter estimation in machine learning which has a low computational cost for the iterations[36, 37].

4 Experimental setup and methods

Under this section we introduce the dataset for this paper, explore our dataset to select the best features for our model construction and describe our methods. We describe the process of setting up our classifiers and the process involved in obtaining our results. The experiments were carried out on a computer running Microsoft Windows 10 home edition with Intel(R) Core(TM) i5 - 7200U CPU @ 2.50GHz and 8GB of RAM.

4.1 The dataset

This paper used data from Kaggle [38], Originally sourced from a mobile money service provider in an African country. It consists of ten(10) columns with their descriptions given below;

1. type is made up of CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER.
2. amount is the amount of the transaction in local currency.
3. nameOrig is the customer who started the transaction.
4. oldbalanceOrg is the initial balance before the transaction
5. newbalanceOrig is the customer's balance after the transaction.
6. nameDest is the recipient ID of the transaction.
7. oldbalanceDest is the initial recipient balance before the transaction.
8. newbalanceDest is the recipient's balance after the transaction.
9. isFraud identifies a fraudulent transaction (1) and non fraudulent (0).
10. isFlaggedFraud flags illegal attempts to transfer more than $200,000$ in a single transaction.

4.2 Data exploration, feature engineering, and selection

The dataset was explored with visualization tools from matplotlib.pyplot and seaborn. The heatmap is reported in Fig. 2. Further analysis showed no significant influence of certain independent variables on the dependent variable in the dataset. These irrelevant variables were subsequently removed from the dataset to enable an efficient model generation.

The dataset was checked further to ascertain the level of independence between the predictors using Spearman's rank correlation coefficient.

The relevant features selected from the previous analysis were further analyzed statistically to determine its suitability for the model development. The results are presented in Table 1.

The results from the Logit Regression Results showed that two of the features extracted for the model development cranked out$p$values of 0.540 and 0.708 which were above the acceptable value of 0.05, they were subsequently dropped. This rigorous selection process legitimizes the appropriateness of the selected features for the model development.

Table 1: Analysis report on the suitability of the independent variables.

Independent Variable	Coefficient	Standard Error	z	p> $|z|$
CASH-OUT	-5.1009	0.024	-211.928	0.00
DEBIT	-16.5023	26.937	-0.613	0.540
PAYMENT	-22.3198	59.545	-0.375	0.708
TRANSFER	-3.2856	0.031	-105.467	0.000
amount	7.555e-05	7.13e-07	-105,987	0.000
oldbalanceOrg	8.539e-05	7.03e-07	118.886	0.000
newbalanceOrig	-8.715e-05	6.96e-07	-125.172	0.000
oldbalanceDest	-7.695e-07	1.78e-08	-43.346	0.000

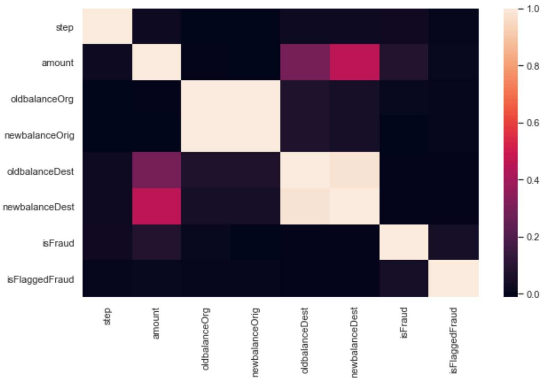

Figure 2: Heatmap representation of the dataset.

4.3 Methods

The methodology was implemented in Python. The approach is illustrated in the flow chart in Fig.3.We began by collecting the dataset. The data was preprocessed to extract the relevant features needed for the model development. The model was then constructed based on the selected features. After the model construction, different evaluation metrics were used to determine the suitability and relevance of the model to the problem at hand. Where appropriate, the model was accepted otherwise the parameters were tuned and the model reconstructed until an adequate one was found.

4.4 Model construction

In order to develop a good model, we analyzed the predictor variable, isFraud, in the paysim1 dataset to determine its distribution. The results showed 8213 for the positive class (1) and 6354407 for the negative class (0) representing 0.1290% and 99.8709% respectively. This is a clear indication of high imbalance in the predictor variables and further manifested in the training dataset (25% of the dataset), which also showed 6186 for the positive class (1) and 4765779 for the negative class (0) representing 0.1296% and 99.8703% respectively. From these values, it is evident that developing a “normal” Logistic Regression model will be biased towards the negative class. However, for the sake of analysis, this paper looked at the result from building a “normal” Logistic Regression which does not take into account the imbalance nature of the dataset as well as the other methods of dealing with data imbalance in machine learning to enable us perform better analysis of the results.

4.5 “Normal" logistic regression

In building the “normal" Logistic regression, we used 25% of the dataset for the training phase. No resampling was performed on the training dataset to see the effect of imbalance on the results. The model produced a score of 99.8747% and a wrong classification score of 0.075%. The rate at which the model was able to detect fraudulent transactions(TPR) was 81.3517% of all the actual fraudulent transactions and the ability to detect fraudulent free transactions(TNR) was 99.8983% of the actual fraudulent free transactions. The rate at which fraudulent free transactions(FPR) was classified as fraud was 79.6743% of all the actual transactions and fraudulent transactions classified as legitimate ones (FNR) was 0.0237% of all the actual fraudulent free transactions. Other classification reports, confusion matrix, root mean square error (RMSE), and AUCROC values are reported in Tables 2,3,4, and 5 respectively. A plot of the receiver operating characteristics curve is also presented in Fig. 6.

4.6 Undersampling

For undersampling, we aimed at removing the tilt towards the negative class of the model by trying to equalize the class lengths of both the majority and the minority in the training dataset. In this method, we reduced the length of the majority class from 4765779 to make it equal to that of the minority class of 6186 which is 25% of the training dataset. We achieved this by removing randomly a number of some of the majority class indices in an attempt to reduce its length to make it equal to the length of the minority class. This method produced an accuracy of 89.5057% a wrong classification score of 10.5637%. The rate at which the model was able to detect fraudulent transactions(TPR) was 97.5099% of all the actual fraudulent transactions and the ability to detect fraudulent free transactions(TNR) was 83.4237% of the actual fraudulent free transactions. The rate at which fraudulent free transactions(FPR) was classified as fraud was 18.9741% of all the actual fraudulent transactions and fraudulent transactions classified as legitimate ones (FNR) was 2.4666% of all the actual fraudulent free transactions. Other classification reports, confusion matrix, RMSE, and AUCROC values are reported in Tables 2,3,4, and 5 respectively. A plot of the receiver operating characteristics curve is also presented in Fig. 6.

4.7 Logistic regression with weight

In this method of the model development, we imposed weights on the class errors which were proportional to the class imbalance. This was achieved by setting the hyper parameter of the logistic regression classifier "weight" to "balanced" from scikit-learn which assigned certain weights to the classes in an effort at balancing the influence both classes have on the classifier. This produced a model score of 96.2240% a wrong classification score of 3.7759%. The rate at which the model was able to detect fraudulent transactions(TPR) was 85.9891% of all the actual fraudulent transactions and the ability to detect fraudulent free transactions(TNR) was 96.2370% of the actual fraudulent free transactions. The rate at which fraudulent free transactions(FPR) was classified as fraud was 2949.13% of all the actual fraudulent transactions and fraudulent transactions classified as legitimate ones (FNR) was 0.0178% of all the actual fraudulent free transactions. Other classification reports, confusion matrix, RMSE, and AUCROC values are reported in Tables 2,3,4, and 5 respectively. A plot of the receiver operating characteristics curve is also presented in Fig. 6.

4.8 Synthetic minority oversampling technique (SMOTE)

In this model, we employed the oversampling technique, SMOTE. SMOTE attempts to increase the size of the minority class by introducing new instances of the minority class in the neighborhood of the minority classes[20]. This method attempts to match the size of the majority and the minority class. The method yielded a length of 4765779 for the positive class (1) and 4765779 for the negative class (0). This model produced a score of 86.2210% a wrong classification score of 13.7789%. The rate at which the model was able to detect fraudulent transactions(TPR) was 98.0266% of all the actual fraudulent transactions and the ability to detect fraudulent free transactions(TNR) was 86.2060% of the actual fraudulent free transactions. The rate at which fraudulent free transactions(FPR) was classified as fraud was 10810.80% of all the actual fraudulent transactions and fraudulent transactions classified as legitimate ones (FNR) was 0.0025% of all the actual fraudulent free transactions. Other classification reports, confusion matrix, RMSE, and AUCROC values are reported in Tables 2,3,4, and 5 respectively. A plot of the receiver operating characteristics curve is also presented in Fig. 6.

4.9 Using smote re-sampling for best parameters (SMOTE RS)

This approach is similar to the method used in (H) and produced 785568 for the positive class (1) and 7855688 for the negative class (0). However, it further employed GridSearchCV[39, 40, 41]to tune the hyper parameters of logistic regression algorithm. This method searched for the best combination of a set of features from a specified grid of possible parameter values. Pipeline was also used to help automate the learning work flows. Pipeline works by enabling a sequence of data to be transformed and correlated together in a model. These two approaches aided in obtaining the best parameters for the SMOTE ratio for optimizing the algorithm. The method obtained 0.01 as the best SMOTE ratio for the model, with the plot of the mean test score against weight reported in Fig. 4. The model produced a score of 98.4941% a wrong classification score of 1.5061%. The rate at which the model was able to detect fraudulent transactions(TPR) was 90.3798% of all the actual fraudulent transactions and the ability to detect fraudulent free transactions(TNR) was 98.5044% of the actual fraudulent free transactions. The rate at which fraudulent free transactions(FPR) was classified as fraud was 1172.0769% of all the actual fruadulent transactions and fraudulent transactions classified as legitimate ones (FNR)was 0.0122% of all the actual fraudulent free transactions. Other classification reports, confusion matrix, RMSE, and AUCROC values are reported in Tables 2,3,4, and 5 respectively. A plot of the receiver operating characteristics curve is also presented in Fig. 6.

4.10 Manual weights tuning (MWT)

Under this approach, we aimed at achieving a trade-off for the harmonic mean by manually tuning the class weights for the false positives and the false negatives. The class size used was 25% of the original dataset, 4765779 for negative class (0) and 6186 for the positive class (1). We achieved this by setting twenty five(25) evenly spaced weight points between 0.01 and 1.0 using GridSearchCV with 5 fold cross validations. 0.8350 was obtained as the best weight parameter for the negative class and 0.1649 for the positive class. These results were then fitted into our logistic regression model. The plot of the mean test score against weight is reported in Fig.5 This model yielded a score of 99.9559% a wrong classification score of 0.0453%. The rate at which the model was able to detect fraudulent transactions(TPR) was 70.1529% of all the actual fraudulent transactions and the ability to detect fraudulent free transactions(TNR) was 99.9926% of the actual fraudulent free transactions. The rate at which fraudulent free transactions(FPR) was classified as fraud was 5.7720% of all the actual fraudulent transactions and fraudulent transactions classified as legitimate ones (FNR) was 0.0380% of all the actual fraudulent free transactions. Other classification reports, confusion matrix, RMSE, and AUCROC values are reported in Tables 2,3,4,

Table 2: Classification report.

Classifier	Precision(%)	Recall(%)	F1 Score(%)
NLR	50.52	81.35	62.33
Undersampling	83.89	97.54	90.20
LRW	2.83	85.98	5.48
SMOTE	0.89	98.02	1.78
SMOTE RS	7.15	90.37	13.26
MWT	92.39	70.15	79.75
ADASYN	0.84	98.86	1.68

and 5 respectively. A plot of the receiver operating characteristics curve is also presented in Fig. 6.

4.11 Adaptive synthetic sampling (ADASYN)

ADASYN is one of the methods for dealing with the problem of class imbalance[21, 42]. ADASYN works by generating additional class samples synthetically for the minority class by using density distribution to automatically determine the number of artificial samples that are to be generated for the minority class[22]. The algorithm produced a score of 85.2557% a wrong classification score of 14.7442%. The rate at which the model was able to detect fraudulent transactions(TPR) was 98.8653% of all the actual fraudulent transactions and the ability to detect fraudulent free transactions(TNR) was 85.2383% of the actual fraudulent free transactions. The rate at which fraudulent free transactions(FPR) was classified as fraud was 11569.1662% of all the actual fraudulent transactions and fraudulent transactions classified as legitimate ones (FNR)was 0.0014% of all the actual fraudulent free transactions. Other classification reports, confusion matrix, RMSE, and AUCROC values are reported in Tables 2,3,4, and 5 respectively. A plot of the receiver operating characteristics curve is also presented in Fig. 6.

5 Performance evaluation, results and discussion

In other to obtain a vivid analysis of our experimental result, we explore a variety of metrics. The following metrics were used in the evaluation of the models; Accuracy, Recall, Precision, F1-Score, AUCROC curve, and RMSE. For a classifier, the confusion matrix output is classified as True Positive(TP), True Negative(TN), False Positive(FP) and False negative(FN). Accuracy is the ratio of the correctly predicted samples to the total of all the samples used in the training set. It is given by equation(8)[43]

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

Recall can be defined as the ratio of true positives to the sum of true positives and false negatives. It is given by equation(9)[43]

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

Precision is defined as the ratio of correctly predicted positive observations to the total predicted positive observations. It is given by equation(10)[43]

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

F1 Score is defined as the weighted average of Precision and Recall. The formula is given by equation(11)[43]

$$F1 - Score = \frac{2 * Recall * precison}{Recall + Precision} \tag{11}$$

AUCROC is defined as the area under the curve of the plot of the true positive rate to the false positive rate [29, 44]. The values range between 0 and 1. As the AUC approaches 1, it is an indication of a better model and a bad model as the value approaches 0. The curve is a plot of True Positive Rate (TPR) Versus False Positive Rate (FPR). It is given by equation(12)[45]

$$AUC = \int_a^b \frac{TP}{TP + FN} d \frac{FP}{TN + FP} \tag{12}$$

Root mean squared error is a square root of the average of squared differences between the observed class and the predicted class. it is given by equation(13)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{13}$$

5.1 Results and discussion

For a model to be considered adequate for classification and used subsequently for prediction, one of the key indicators is the evaluation of the TPR, TNR, FPR, and FNR. The TPR and TNR should be high as possible whiles the FPR and FNR needs to be as low as possible. From Figure7, which is a Bar chart of Classifiers and their respective TPR,

Table 3: Confusion Matrix for the Models.

		Predicted Class	
Classifier	Actual Class	0	1
NLR	0	1587013	1615
	1	378	1649
Undersampling	0	1691	381
	1	50	1958
LRW	0	1528849	59779
	1	284	1743
SMOTE	0	1369493	219135
	1	40	1987
SMOTE RS	0	1564870	23758
	1	195	1832
MWT	0	1588511	117
	1	605	1422
ADASYN	0	1354121	234507
	1	23	2004

Table 4: Model Error.

Classifier	RMSE
NLR	0.0353
Undersampling	0.3239
LRW	0.1943
SMOTE	0.3711
SMOTE RS	0.1277
MWT	0.0213
ADASYN	0.3839

Table 5: AUC Values for the Models.

Classifier	AUC value
NLR	0.9642
Undersampling	0.9763
LRW	0.9740
SMOTE	0.9813
SMOTE RS	0.9813
MWT	0.9627
ADASYN	9826

TNR, FPR, and FNR of the models used in our experiments, all the seven(7) models produced TPR and TNR values that exceed 70% in their respective domains. The FNR for all the experiments also had values that were below 3% of all fraudulent transactions classified as legitimate ones. The rate at which legitimate transactions were classified as fraudulent ones(FPR) was not encouraging enough as five(5) out of the seven(7) models had values of over 79% of of all the actual fraudulent transactions with the maximum reaching 11569.1662%. These models cannot be implemented since it will cause a lot of anxiety and frustrations for legitimate users which can lead to customer churn. This leaves only two(2) models left for consideration; Undersampling and MWT. Undersampling had 18.9741% for FPR and 2.4666% for FNR while MWT had 5.7720% for FPR and 0.0380 for FNR. We therefore accept MWT as the best model under this evaluation.

We proceeded to analyze our model based on the F1 Score. Four out of seven models achieved F1 Score values of below 6% which are too low to be considered for inclusion in our model development. NLR, Undersampling and MWT are therefore the models left for consideration. Undersampling acheived the highest score of 90.20% followed by MWT 79.75% and NLR 62.33%.

Our next evaluation metric was the AUC ROC values. The AUC values produced by all the seven(7) models exceeded 0.9 which makes them all very good models per this evaluation.

We now consider the RMSE which were in the range of 0.0213 for MWT to 0.3839 for ADASYN.

In this experiments two models came top as good models; MWT and Undersampling. Manual Weights Tuning achieved a model score of 99.9559% F1-score with a value of 79.75%, the lowest in the RMSE and an AUC value of 0.9627. Undersampling recorded good results, having the best F1-score of 90.20% and an accuracy of 89.5057%. Undersampling, however, discarded a large chunk of the dataset, utilizing only 4080 for testing compared to 1 590 655 for the other models, making it inappropriate for our model. We therefore consider MWT as the best model in our experiments.

In comparison with other works, MWT achieved a superior performance compared with [46] which obtained 92.74% using C4.5. [47] who obtained 97% to 98% using cased based reasoning(CBR). MWT also performed better as compared to a similar work by [23] who experimented with a hybrid technique for undersampling and oversampling achieving 97.92% for Naive Bayes, 97.69% for k-nearest neighbor and 54.86% for LR

The other models performed poorly with each obtaining harmonic means of less than 10%. Even though

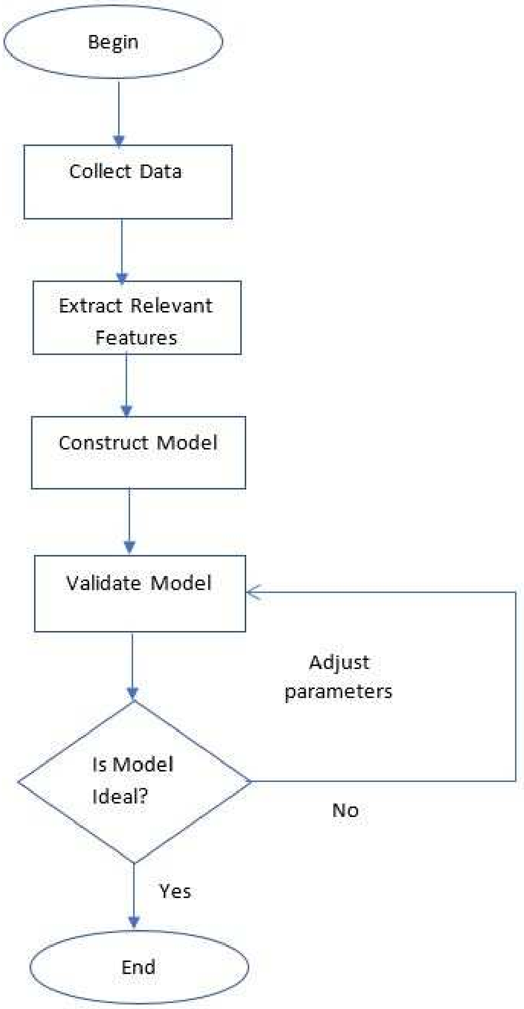

Figure 3: A flowchart of the model development.

other parameters were used, the F1-score was one of the key metrics since it is normally used to evaluate prediction(classification) algorithms because of it's ability to balance the effect on recall and precision[48].

6 Conclusion

In this article, we looked at different approaches on how to classify and predict fraud cases in MMTs with keen interest in its associated class imbalance problem. We have shown the effects different resampling techniques have on our prediction (classification) results. We further indicated this by looking at different evaluation metrics. Our best model for this experiments was the manual tuning of the class weights for the false positives and the false negatives. This was aimed at achieving a trade off for the F1-score. We also demonstrated the practicality of our work using logistic regression.

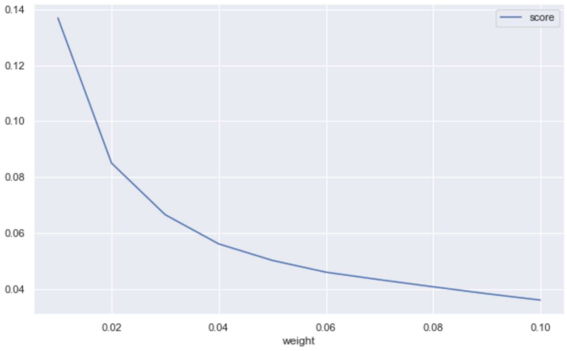

Figure 4: Plot of the mean test score against weight for SMOTE RS.

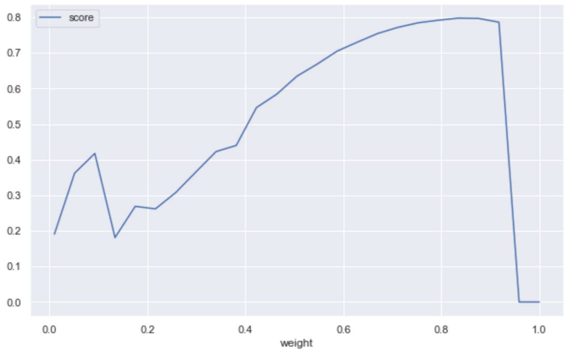

Figure 5: Plot of the mean test score against weight for MWT.

Acknowledgment

This work was supported in part by the Frontier Science and Technology Innovation Projects of National Key R&D Program (No.2019QY1405), the National Natural Science Foundation of China (No. 61672135), the Sichuan Science-Technology Support Plan Program (No. 2018GZ0236 and No. 2017FZ0004), the Fundamental Research Funds for the Central Universities (No.2672018ZYGX2018J057), and the CERNET Innovation Project(No. NGII20180404).

References

[1] S. Yu and S. Ibtasam, "A qualitative exploration of mobile money in ghana," in *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*. ACM, 2018, p. 21. [Online]. Available: https://doi.org/10.1145/3209811.3209863

[2] M. Zhdanova, J. Repp, R. Rieke, C. Gaber, and B. Hemery, "No smurfs: Revealing fraud chains in mobile money transfers," in *2014 Ninth International Conference on Availability, Reliability and Security*. IEEE, 2014, pp. 11–20. [Online]. Available: https://doi.org/10.1109/ares.2014.10

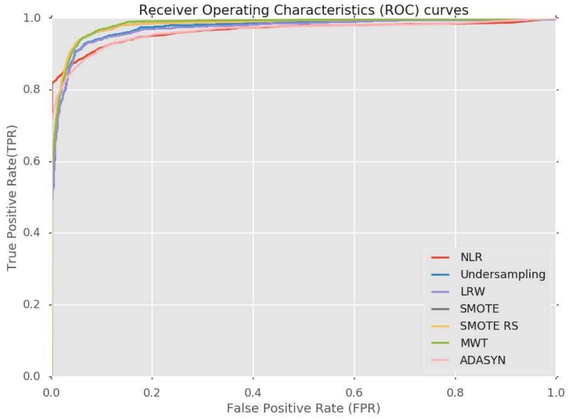

Figure 6: Plots of the ROC curves.

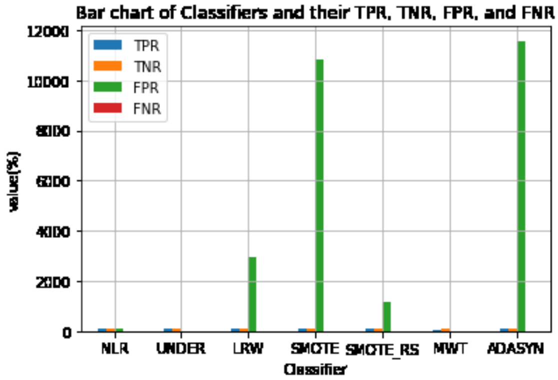

Figure 7: Bar chart of Classifiers and their TPR, TNR, FPR, and FNR.

[3] A. Demirguc-Kunt, L. Klapper, D. Singer, and P. Van Oudheusden, *The global findex database 2014: Measuring financial inclusion around the world*. The World Bank, 2015.

[4] http://www.gsma.com/mobilefordevelopment/wp-content/uploads/2017/03/GSMA_State-of-the-Industry-Report-on-Mobile-Money_2016.pdf, Accessed: 2019-05-10.

[5] I. Akomea-Frimpong, C. Andoh, A. Akomea-Frimpong, and Y. Dwomoh-Okudzeto, "Control of fraud on mobile money services in ghana: an exploratory study," *Journal of Money Laundering Control*, 2019. [Online]. Available: https://doi.org/10.1108/jmlc-03-2018-0023

[6] S. A. AlSaif and A. Hidri, "Impact of data balancing during training for best predictions," vol. 45, no. 2, Jun. 2021. [Online]. Available: https://doi.org/10.31449/inf.v45i2.3479

[7] A. A. Abaker and F. A. Saeed, "A comparative analysis of machine learning algorithms to build a predictive model for detecting diabetes complications," vol. 45, no. 1, Mar. 2021. [Online]. Available: https://doi.org/10.31449/inf.v45i1.3111

[8] H. Barros and M. Silveira, "Atlas based sparse logistic regression for alzheimer's disease classification," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2017, pp. 501–504. [Online]. Available: https://doi.org/10.1109/embc.2017.8036871

[9] G. Harshvardhan, N. Venkateswaran, and N. Padmapriya, "Assessment of glaucoma with ocular thermal images using glcm techniques and logistic regression classifier," in *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*. IEEE, 2016, pp. 1534–1537. [Online]. Available: https://doi.org/10.1109/wispnet.2016.7566393

[10] L. Li, X. Wang, X. Du, Y. Liu, C. Liu, C. Qin, and Y. Li, "Classification of heart sound signals with bp neural network and logistic regression," in *2017 Chinese Automation Congress (CAC)*. IEEE, 2017, pp. 7380–7383. [Online]. Available: https://doi.org/10.1109/cac.2017.8244111

[11] W. Pramesti, I. Damayanti, and D. A. Asfani, "Stator fault identification analysis in induction motor using multinomial logistic regression," in *2016 International Seminar on Intelligent Technology and Its Applications (ISITIA)*. IEEE, 2016, pp. 439–442. [Online]. Available: https://doi.org/10.1109/isitia.2016.7828700

[12] J. Gao, S. Feng, Q. Huang, Z. Zhang, R. Luo, and Y. Teng, "A study of logistic regression-based discrimination method of false overcurrent alarm of 500kv high-voltage shunt reactor," in *2018 International Conference on Smart Grid and Clean Energy Technologies (ICSGCE)*. IEEE, 2018, pp. 218–222. [Online]. Available: https://doi.org/10.1109/icsgce.2018.8556698

[13] D. Prasetio *et al.*, "Predicting football match results with logistic regression," in *2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*. IEEE, 2016, pp. 1–5. [Online]. Available: https://doi.org/10.1109/icaicta.2016.7803111

[14] A. Vaidya, "Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval," in *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 2017, pp. 1–6. [Online]. Available: https://doi.org/10.1109/icccnt.2017.8203946

[15] E. K. Ampomah, G. Nyame, Z. Qin, P. C. Addo, E. O. Gyamfi, and M. Gyan, "Stock market prediction with

gaussian naïve bayes machine learning algorithm," *Informatica*, vol. 45, no. 2, 2021. [Online]. Available: https://doi.org/10.31449/inf.v45i2.3407

[16] E. K. Ampomah, Z. Qin, G. Nyame, and F. E. Botchey, "Stock market decision support modeling with tree-based adaboost ensemble machine learning models," *Informatica*, vol. 44, no. 4, 2021. [Online]. Available: https://doi.org/10.31449/inf.v44i4.3159

[17] T. Liu and L. Zhang, "Application of logistic regression in web vulnerability scanning," in *2018 International Conference on Sensor Networks and Signal Processing (SNSP)*. IEEE, 2018, pp. 486–490. [Online]. Available: https://doi.org/10.1109/snsp.2018.00097

[18] R. R. Popat and J. Chaudhary, "A survey on credit card fraud detection using machine learning," in *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE, 2018, pp. 1120–1125. [Online]. Available: https://doi.org/10.1109/icoei.2018.8553963

[19] J. L. Perols, R. M. Bowen, C. Zimmermann, and B. Samba, "Finding needles in a haystack: Using data analytics to improve fraud prediction," *The Accounting Review*, vol. 92, no. 2, pp. 221–245, 2016. [Online]. Available: https://doi.org/10.2308/accr-51562

[20] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla, "Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," *Journal of artificial intelligence research*, vol. 61, pp. 863–905, 2018. [Online]. Available: https://doi.org/10.1613/jair.1.11192

[21] C. S. Lai, Y. Tao, F. Xu, W. W. Ng, Y. Jia, H. Yuan, C. Huang, L. L. Lai, Z. Xu, and G. Locatelli, "A robust correlation analysis framework for imbalanced and dichotomous data with uncertainty," *Information Sciences*, vol. 470, pp. 58–77, 2019. [Online]. Available: https://doi.org/10.1016/j.ins.2018.08.017

[22] A. Gosain and S. Sardana, "Handling class imbalance problem using oversampling techniques: A review," in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2017, pp. 79–85. [Online]. Available: https://doi.org/10.1109/icacci.2017.8125820

[23] J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis," in *2017 International Conference on Computing Networking and Informatics (ICCNI)*. IEEE, 2017, pp. 1–9. [Online]. Available: https://doi.org/10.1109/iccni.2017.8123782

[24] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, and A. Anderla, "Credit card fraud detection-machine learning methods," in *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*. IEEE, 2019, pp. 1–5. [Online]. Available: https://doi.org/10.1109/infoteh.2019.8717766

[25] G. Rushin, C. Stancil, M. Sun, S. Adams, and P. Beling, "Horse race analysis in credit card fraud—deep learning, logistic regression, and gradient boosted tree," in *2017 Systems and Information Engineering Design Symposium (SIEDS)*. IEEE, 2017, pp. 117–121. [Online]. Available: https://doi.org/10.1109/sieds.2017.7937700

[26] S. Maes, K. Tuyls, and B. Vanschoenwinkel, "Machine learning techniques for fraud detection," Ph.D. dissertation, Master's thesis, Vrije Universiteit Brussel, Brussels, 2001.

[27] K. Fu, D. Cheng, Y. Tu, and L. Zhang, "Credit card fraud detection using convolutional neural networks," in *International Conference on Neural Information Processing*. Springer, 2016, pp. 483–490. [Online]. Available: https://doi.org/10.1007/978-3-319-46675-0_53

[28] M. Ahmed, A. N. Mahmood, and J. Hu, "A survey of network anomaly detection techniques," *Journal of Network and Computer Applications*, vol. 60, pp. 19–31, 2016. [Online]. Available: https://doi.org/10.1016/j.jnca.2015.11.016

[29] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006. [Online]. Available: https://doi.org/10.1016/j.patrec.2005.10.010

[30] M. Hossin and M. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, p. 1, 2015. [Online]. Available: https://doi.org/10.5121/ijdkp.2015.5201

[31] J. Ran, G. Zhang, T. Zheng, and W. Wang, "Logistic regression analysis on learning behavior and learning effect based on spoc data," in *2018 13th International Conference on Computer Science & Education (ICCSE)*. IEEE, 2018, pp. 1–5. [Online]. Available: https://doi.org/10.1109/iccse.2018.8468834

[32] X. Wang, L. Song, L. Sun, and H. Gao, "Nonparametric estimation of the roc curve based on the bernstein polynomial," *Journal of Statistical Planning and Inference*, vol. 203, pp. 39–56, 2019. [Online]. Available: https://doi.org/10.1016/j.jspi.2019.02.004

[33] R. Zhu and S. Ghosal, "Bayesian semiparametric roc surface estimation under verification bias,"

Computational Statistics & Data Analysis, vol. 133, pp. 40–52, 2019. [Online]. Available: https://doi.org/10.1016/j.csda.2018.09.003

[34] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of research and development*, vol. 3, no. 3, pp. 210–229, 1959.

[35] T. Mitchell, *Machine Learning*. United States: McGraw Hill, 1997.

[36] A. Mokhtari and A. Ribeiro, "Global convergence of online limited memory bfgs," *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 3151–3181, 2015.

[37] M. Ahookhosh, K. Amini, M. Kimiaei, and M. Peyghami, "A limited memory adaptive trust-region approach for large-scale unconstrained optimization," *Bulletin of the Iranian Mathematical Society*, vol. 42, no. 4, pp. 819–837, 2016.

[38] E. Lopez-Rojas, A. Elmir, and S. Axelsson, "Paysim: A financial mobile money simulator for fraud deection," in *28th European Modeling and Simulation Symposium, EMSS, Larnaca*. Dime University of Genoa, 2016, pp. 249–255.

[39] 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," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[40] R. Garreta and G. Moncecchi, *Learning scikit-learn: machine learning in python*. Packt Publishing Ltd, 2013.

[41] Y. Shuai, Y. Zheng, and H. Huang, "Hybrid software obsolescence evaluation model based on pca-svm-gridsearchcv," in *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*. IEEE, 2018, pp. 449–453. [Online]. Available: https://doi.org/10.1109/ICSESS.2018.8663753

[42] M. Koziarski, B. Krawczyk, and M. Woźniak, "Radial-based oversampling for noisy imbalanced data classification," *Neurocomputing*, vol. 343, pp. 19–33, 2019. [Online]. Available: https://doi.org/10.1016/j.neucom.2018.04.089

[43] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *Ieee Access*, vol. 5, pp. 8869–8879, 2017. [Online]. Available: https://doi.org/10.1109/ACCESS.2017.2694446

[44] F. Isinkaye, Y. Folajimi, and B. Ojokoh, "Recommendation systems: Principles, methods and evaluation," *Egyptian Informatics Journal*, vol. 16, no. 3, pp. 261–273, 2015. [Online]. Available: https://doi.org/10.1016/j.eij.2015.06.005

[45] R. Vinayakumar, M. Alazab, K. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, "Deep learning approach for intelligent intrusion detection system," *IEEE Access*, vol. 7, pp. 41 525–41 550, 2019.

[46] A. Husejinovic, "Credit card fraud detection using naive bayesian and c4. 5 decision tree classifiers," *Husejinovic, A.(2020). Credit card fraud detection using naive Bayesian and C*, vol. 4, pp. 1–5, 2020.

[47] A. Adedoyin, S. Kapetanakis, G. Samakovitis, and M. Petridis, "Predicting fraud in mobile money transfer using case-based reasoning," in *International Conference on Innovative Techniques and Applications of Artificial Intelligence*. Springer, 2017, pp. 325–337. [Online]. Available: https://doi.org/10.1007/978-3-319-71078-5_28

[48] T. Liu, S. Wang, S. Wu, J. Ma, and Y. Lu, "Predication of wireless communication failure in grid metering automation system based on logistic regression model," in *2014 China International Conference on Electricity Distribution (CICED)*. IEEE, 2014, pp. 894–897.