A Fraud Detection Framework using Machine Learning Approach

Aderonke Thompson
Department of Cyber Security
Federal University of Technology
Akure, Nigeria
afthompson@futa.edu.ng

Oghenerukvwe Oyinloye
Department of Computer Science
Ekiti State University of Technology
Ado Ekiti, Nigeria
oeoyinloye@eksu.edu.ng

Leon Aborisade
Department of Computer Science
Federal University of Technology
Akure, Nigeria
aborisadelo@futa.edu.ng

Esther Odeniyi
Department of Cyber Security
Federal University of Technology
Akure, Nigeria
eaodeniyi@futa.edu.ng

***Abstract*— Credit card fraud describes cases in which a threat actor gains unauthorized access in order to obtain money or property. The importance of machine learning and Data Science cannot be over emphasized. This work develops an efficient fraud detection framework using non-rule-based approach of Multi-layer perceptron. It correctly predicts and detects frauds on a given financial transaction dataset. The algorithm on the datasets evaluates its effectiveness vis-à-vis frauds detection in bank transactions. The results are compared and evaluated using various evaluation metrics.**

***Keywords- fraud; credit cards; Multi-layer perceptron*;**

I. Introduction

Recent information technology (IT) proliferation deployed in major financial services by Nigerian banking institutions has led to an increase in threats posed to these systems. Debit/Credit cards are one of the most common payment methods used over the Internet. It was asserted that financial fraud can be viewed as an act intended for deception involving financial transactions for personal gain purpose [1]. Fraudsters have it easier as most transactions do not require the presence of a bank account/card holder; stealing relevant customers details or perform identity theft by posing as the customer at point of payments is all that is vital to perpetrating their acts. This includes phishing and unsuspecting customers, redirection to malicious websites with a hidden act of harvesting customers' banking details and information. Credit card fraud is equally viewed as a type of theft and fraud done using a payment card, as a fraudulent fund source in a transaction. Some security issues are mostly faced by banks everywhere, but the prevention of card fraud attracts high priority, and this is set to grow with the exponential rate of Internet awareness and transactions. Increase in online purchases has made criminals take advantage of various weak authentication checks to commit credit card fraud [2].

Models provide a way to mitigate these occurrences, protect clients' transactions and play an essential role in payment service providers' profitability and sustainability. All the aforementioned can be achieved using a fraud detection system (FDS). FDS is computational analysis fraud detection techniques via fraud identification or anomaly transactions in swift and proven techniques of machine learning as presented in [3]. Modeling of past credit card transactions has to do with detecting fraudulent transactions via the existing knowledge fraud. This model is then used to identify whether a new transaction is fraudulent or not in the two major existing fraud methods of physical and virtual frauds. Physical fraud is done by stealing a card and using it for the payment or purchasing while virtual fraud is committed by using someone's card details through the internet for transactions. Further classification of credit card fraud is given in Figure.1. Section I deals with the introduction of various acts of fraud. A guide to available credit card fraud is presented in Section II, while Section III gives a detail of related study in the fraud detection domain. In Section IV, multilayer perceptron methodology approach to fraud detection is extensively discussed. Implementation of a feed forward Artificial Neural Network for the machine learning approach is presented in Section V in addition to Section VI which further shows the implementation with various parameters. Observations from the proposed model and evaluation is given in Section VII with the performance of the Logistic regression study based on the same dataset.

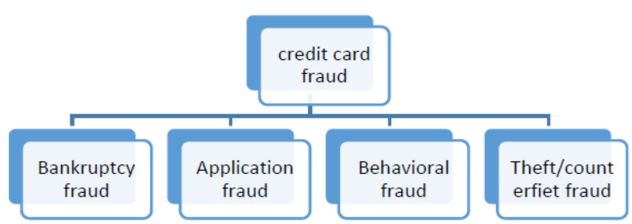

Figure 1. Classification of Credit Card Fraud

II. Major Methods Used to Mitigate Credit Card

There are basically two major forms of mitigating credit card fraud, it could be in the preventive or detective mode. The preventive mode involves blocking fraudulent transaction at the point of transaction. Such as passwords, pin and blocked cards; while the detective mode identifies successful fraud transaction through predictive models with machine learning approach.

Traditionally, fraud resolution process usually involves: fraud detection, investigation, confirmation, and prevention. Therefore, a self-learning computer program automates the above processes using various methods. Signature based detection method detects fraud traces through the signature technology using known patterns or byte sequence, it is efficient for known frauds. However, fraudsters have continued to manipulate the system by finding creative ways to beat signature strings. The anomaly detection method comes with the ability to detect both known and novel frauds; although, this method is limited by false positive error, that is, previously unknown legitimate transactions. Consequently, this paper exploits machine learning (see Section IV) to detect fraudulent activities as well as measuring its performance.

III. LITERATURE REVIEW

Financial fraud had been a major challenge for corporate organizations, government and most specifically businesses that utilize information technology. Financial fraud is defined as an intentional act of deception involving financial transactions for personal purpose gain. Another definition for financial fraud is “to take advantage over another by false representations” which include “surprise, trickery, cunning and unfair ways through which another is cheated” [1]. Globally, fraud costs some financial industry approximately $80 billion annually while the United States’ credit and debit card issuers alone lost $2.4 billion.

The financial fraud occurrence in any organization undermines both the effort and prospects. Financial fraud brings about losses owing to theft, distrust in transaction, and litigation. These losses owing to fraud are grossly detrimental to institutions in which they occur. As advances in cloud technology plums and cyber-security measures is not commensurate,, there exists high possibility of financial fraud bound to threaten businesses worldwide. Detection of financial fraud had not come so easy; it is mostly at a high cost and time. The cost of financial fraud reported is about $1 million per incident, occupational fraud costs $150 to $200,000 per incident while losses due to fraud costs an average of 5% of gross profit and take around 24 to 36 months to discover - usually via a tip (40%), by accident (20%), or during an audit (10%). Some motivations for committing financial fraud has been reported and identified by senior management to be most responsible for most fraud [4].

The authors in [4] argued that meeting external forecasts emerged as the primary motivation and it was conceptualized that three elements common among all fraud is called the fraud triangle. These elements include a perceived pressure, a perceived opportunity, and a rationalization of the fraud act. in addition to the trio, is motivation for need, greed and addictions (or vices). This is with the assertion that the motivation for greed in turn feeds the motivation for vices. Capping it all, these motivations become a vicious cycle leading to fraud. thus, financial fraud is categorized mainly into three areas: bank fraud, corporate fraud and insurance fraud. Bank fraud is subdivided into credit card fraud, mortgage fraud and money laundering fraud [5].

Fraud modelling is one important tool in addressing financial fraud. It expands in importance as corporate organizations and government determine which type of models to use and continuous update in order to protect against evolving threats. In the past, traditional fraud models are used to automatically detect unauthorized transactions such as determining when a card has been used without the owner’s consent. Most card issuers use fraud models to identify fraudulent card usage in order to maintain the integrity and security of their network as it is core to earning trust in online business world. However, diverse range of payment services offered by organizations and businesses to clients also presents higher opportunities for fraud occurrence. Consequently, fraud models provide a way to mitigate these occurrences, protect clients’ transactions and play an essential role in payment service providers’ profitability and sustainability with attributes of a given transaction as variables used in fraud models. Thereafter, it classifies or attempts to label the transaction fraudulent or legitimate (see Section V-VII). Some extensive models label the type or category of fraud. Some of the common attributes used by fraud models include: Merchant (the business charging the transaction), transaction location, amount, type (online or offline), volume, account history, transaction history, and so on, depending on the amount of attribute information captured in a transaction. The five basic fields, which describe type, time (hours, minutes), location, amount, and date (week days) of a transaction were used in the fraud model. While 16 significant ratios out of 29 financial ratios were used in detection of fraud in the financial statements of banks which were categorized into asset quality ratios, earnings and profitability ratios, liquidity/solvency ratios, long term solvency/leverage ratio, capital adequacy ratio, cash flow analysis and trends. These fraud models utilized 29 variables of which 24 are financial variables while 5 are non-financial variables as it proved that model tools based on financial numbers, linguistic behaviour, and non-verbal vocal cues have each demonstrated the potential for detecting financial fraud. Fifty-one (51) financial ratios were utilized in detecting fraud in financial statements by means of financial ratios [6].

Notable fraud detection models are mainly categorized as rule-based models and algorithmic (or machine learning) models. Rule-based models are collection of rules used to detect fraudulent transactions with a single rule containing as a set of conditions that, when present, labels a transaction either as fraudulent or not. Rule-based models are made up of an expert knowledge base. In addition, new rules evolve from time to time because of inference action on streams of time changing data. However, one major limitation of rule-based fraud models is time complexity in handling big data. Algorithmic models make use of machine-learning methods to classify a transaction as either fraudulent or legitimate. Algorithmic models are more complex than rule-based models; this is dependent on the type of algorithm used. These models are computationally complex than rule-based models but achieve high performance. They are far better at detecting complex relationships between variables than the rule-based models. Machine-learning methods also require a pre-requisite of having many variables to implement and ensure learning. Therefore, when there is limited number of variables usage, the benefit of algorithmic methods over rule-based models is diminished.

The review on financial accounting fraud detection based on data mining techniques was motivated by the idea that the failure of internal auditing system of the organization in identifying the accounting frauds has led to the use of specialized procedures to detect financial accounting fraud. The findings of this review showed that data mining techniques such as logistic models, neural networks, Bayesian belief network, and decision trees have been applied most extensively to provide primary solutions to the problems inherent in the detection and classification of fraudulent data. In [6], financial fraud detection using vocal, linguistic and financial cues is presented and observed that these methods for automating financial fraud detection (FFD) have mainly relied on financial statistics; although, some recent studies have suggested that linguistic or vocal cues may also be useful indicators of deception. The hypothesis investigated in the study is that an improved tool (based on financial numbers, linguistic behaviour, and non-verbal vocal cues) could be developed if specific attributes from these feature categories were analyzed concurrently. A set of 1,572 public company quarterly earnings conference call audio file samples was used in the study. The authors re-affirmed that earnings from conference calls are ideal for investigation because they involved corporate executives publicly discussing financial information, thereby simultaneously providing financial, linguistic and vocal cues. The study proved that tools based on financial numbers, linguistic behaviour, and non-verbal vocal cues have each demonstrated the potential for detecting financial fraud. However, it is quite tasking (and computationally intensive) to concurrently source and compute large amount of vocal and linguistic data [7].

In another study, a difference between precision-recall and Receiver Operator Characteristic (ROC) curves for evaluating the performance of credit card fraud detection models was motivated by the need to solve the problem of fraudulent transactions detection with use of machine learning for legitimate or fraudulent the credit card transactions classification. In order to solve this problem, the precision-recall curves are described as an approach. Weighted logistic regression is used as an algorithm level technique and random under-sampling is proposed as data-level technique to build credit card fraud classifier. Performance evaluation of these approaches adopted the ROC curves, which showed the variance of the number of correctly classified positive examples with the number of incorrectly classified negative examples. However, ROC curves present an overly optimistic performance view. It established that precision-recall curves have more advantages than ROC curves in dealing with credit card fraud detection. Nevertheless, the study was limited by inability to find the best solution to the problem of imbalanced data in the dataset [8].

In the same vein, a study on “Combatting Financial Fraud: A Co-evolutionary Anomaly Detection Approach” evolved around the motivation of the major difficulty in anomaly detection which lies in discovering boundaries between normal and anomalous behaviour. The objective was to present a co-evolutionary algorithm which tackles the anomaly detection problem and discover the boundary between normal and abnormal behaviour. The co-evolutionary algorithm was used to provide a competitive interaction between different populations which minimize detection errors and the adaptive evolutionary environment accelerated by the process of finding good solution. The authors implemented the algorithm using anonymized transactional data from a real financial institution. The data set contains two-year Automated Bank Machine (ABM) and Point of Sale (POS) fraud-free transaction history. The research has contributed to knowledge by using concept of evolution to detect anomalies in fraudulent transactions only it was not applied to realistic data [9].

IV. METHODOLOGY

The study deploys multilayer perceptron approach to detect fraud using financial datasets. Each transaction by a customer on card contains the transaction API, which is stripped into attributes. The attributes (model variables) from the API include; Source IP address, Destination IP address, Card pan, Location of transaction, Item bought, Unit of items bought, Amount of transaction and the Date and Time of transaction. The model architectural design is depicted in Figure. 2. The Architecture is divided into 3 major parts, namely:

i. Data preprocessing & Feature Selection
ii. Data Training & Learning
iii. Classification

Financial credit card datasets were selected (Dataset 1 and Dataset 2) were obtained from “Kaggle Data Repository” which are publicly available containing anonymized real-life credit card transactions with an evident presence of fraudulent cases. Dataset 1 was obtained from Kaggle Data Repository, and contains anonymized data to protect user’s vital information. Data was from Credit Card Transactions for users in Europe in 2013. It has 284,808

Copyright (c) IARIA, 2019. ISBN: 978-1-61208-743-6

entries. It has 31 attributes with class labels The Dataset 1 sample is shown in Table 1.

Dataset 2 contains anonymized data to protect users' vital information, Data contains credit card transactions. It has 151,113 entries. It has 11 attributes with class labels, partitioned into testing set and training set. Training set contained 105,778 records and testing set had 45,335 records. Sample records of the Dataset 2 are shown in Table II

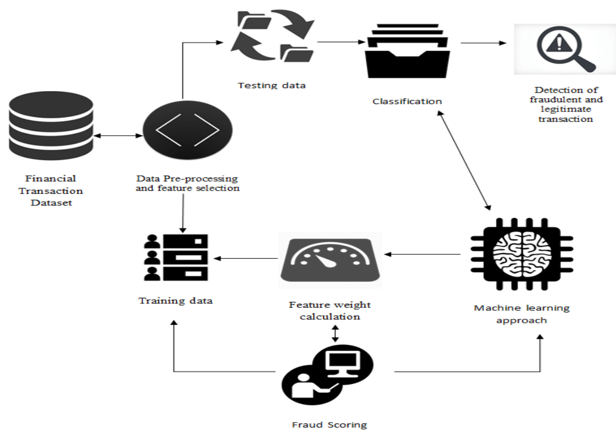

Figure 2. Architectural design of the model

TABLE I. SAMPLE OF DATASET 1

1	Time	V1	V2	V3	V4	V5	V6	V7	V8
2	0	-1.35981	-0.07278	2.536347	1.378155	-0.33832	0.462388	0.239599	0.098698
3	0	1.191857	0.266151	0.16648	0.448154	0.060018	-0.08236	-0.0788	0.085102
4	1	-1.35835	-1.34016	1.773209	0.37978	-0.5032	1.800499	0.791461	0.247676
5	1	-0.96627	-0.18523	1.792993	-0.86329	-0.01031	1.247203	0.237609	0.377436
6	2	-1.15823	0.877737	1.548718	0.403034	-0.40719	0.095921	0.592941	-0.27053
7	2	-0.42597	0.960523	1.141109	-0.16825	0.420987	-0.02973	0.476201	0.260314
8	4	1.229658	0.141004	0.045371	1.202613	0.191881	0.272708	-0.00516	0.081213
9	7	-0.64427	1.417964	1.07438	-0.4922	0.948934	0.428118	1.120631	-3.80786
10	7	-0.89429	0.286157	-0.11319	-0.27153	2.669599	3.721818	0.370145	0.851084
11	9	-0.33826	1.119593	1.044367	-0.22219	0.499361	-0.24676	0.651583	0.069539
12	10	1.449044	-1.17634	0.91386	-1.37567	-1.97138	-0.62915	-1.42324	0.048456
13	10	0.384978	0.616109	-0.8743	-0.09402	2.924584	3.317027	0.470455	0.538247
14	10	1.249999	-1.22164	0.38393	-1.2349	-1.48542	-0.75323	-0.6894	-0.22749
15	11	1.069374	0.287722	0.828613	2.71252	-0.1784	0.337544	-0.09672	0.115982
16	12	-2.79185	-0.32777	1.64175	1.767473	-0.13659	0.807596	-0.42291	-1.90711
17	12	-0.75242	0.345485	2.057323	-1.46864	-1.15839	-0.07785	-0.60858	0.003603
18	12	1.103215	-0.0403	1.267332	1.289091	-0.736	0.288069	-0.58606	0.18938
19	13	-0.43691	0.918966	0.924591	-0.72722	0.915679	-0.12787	0.707642	0.087962
20	14	-5.40126	-5.45015	1.186305	1.736239	3.049106	-1.76341	-1.55974	0.160842

creditcard

The data pre-processing and preparation was carried out on the raw financial dataset to remove outliers using max-min normalization technique. As shown in equation (1)

$$Normalized\_Value = \frac{(f_{value} - f_{\min})}{(f_{\max} - f_{\min})} \tag{1}$$

where f_{value}, is the feature value to be normalized, f_{min} is the minimum feature value and f_{max} is the maximum feature value respectively.

Feature selection was performed by computing feature importance. This is done using Information gain calculation. Thus, given a set of financial transaction dataset S_c

$$E(F) = \sum_{j=1}^{c} \frac{S1_j + \ldots + Sc_j}{S} * I(si_j, \ldots, sc_j) \tag{2}$$

where (I = information, S = total number of financial transaction data instances, c = total classes (i.e. fraudulent and legitimate classes, F = Features)

The information gain, G(F) is defined as:

$$G(F) = I(s_1, s_2, \ldots, s_c) - E(F) \tag{3}$$

Features with high information gain are selected for model development while the others are removed.

TABLE II: SAMPLE OF DATASET 2

user_id	signup_time	purchase_time	purchase_	device_id	source	browser	sex	age	ip_address	class
22058	2/24/2015 22:55	4/18/2015 2:47	34	QVPSPJUC	SEO	Chrome	M	39	732758368.8	0
333320	6/7/2015 20:39	6/8/2015 1:38	16	EOGFQPIZ	Ads	Chrome	F	53	350311387.9	0
1359	1/1/2015 18:52	1/1/2015 18:52	15	YSSKYOSJH	SEO	Opera	M	53	2621473820	1
150084	4/28/2015 21:13	5/4/2015 13:54	44	ATGTXKYK	SEO	Safari	M	41	3840542444	0
221365	7/21/2015 7:09	9/9/2015 18:40	39	NAUITBZF.	Ads	Safari	M	45	415583117.5	0
159135	5/21/2015 6:03	7/9/2015 8:05	42	ALEYXFXIN	Ads	Chrome	M	18	2809315200	0
50116	8/1/2015 22:40	8/27/2015 3:37	11	IWKVZHJO	Ads	Chrome	F	19	3987484329	0
360585	4/6/2015 7:35	5/25/2015 17:21	27	HPUCUYLN	Ads	Opera	M	34	1692458728	0
159045	4/21/2015 23:38	6/2/2015 14:01	30	ILXYDOZIH	SEO	IE	F	43	3719094257	0
182338	1/25/2015 17:49	3/23/2015 23:05	62	NRFFPPHZ	Ads	IE	M	31	341674739.6	0
199700	7/11/2015 18:26	10/28/2015 21:59	13	TEPSJVVX(	Ads	Safari	F	35	1819008578	0
73884	5/29/2015 16:22	6/16/2015 5:45	58	ZTZZJUCRD	Direct	Chrome	M	32	4038284553	0
79203	6/16/2015 21:19	6/21/2015 3:29	18	IBPNKSMC	SEO	Safari	M	33	4161540927	0
299320	3/3/2015 19:17	4/5/2015 12:32	50	RMKQNVE	Direct	Safari	M	38	3178510015	0
82931	2/16/2015 2:50	4/16/2015 0:56	15	XKIFNYUZI	SEO	IE	M	24	4203487754	0
31383	2/1/2015 1:06	3/24/2015 10:17	58	UNUAVQX	SEO	Safari	F	24	995732779	0
78986	5/15/2015 3:52	8/11/2015 2:29	57	TGHVAWE	SEO	FireFox	M	23	3503883392	0
119824	3/20/2015 0:31	4/5/2015 7:31	55	WFIIFCPIC	Ads	Safari	M	38	131423.789	0
357386	2/3/2015 0:48	3/24/2015 18:27	40	NWSVDOH	Ads	FireFox	M	24	3037372279	0
289172	7/17/2015 5:48	11/12/2015 22:08	46	KFZGQIWD	Direct	FireFox	F	53	1044590098	0

V. MULTI LAYER PERCEPTRON (MLP)

The implementation is a feed-forward artificial neural networks; MLP consists of the input layer, output layer, and one or more hidden layers. Each layer of MLP includes one or more neurons directionally linked with the neurons from the previous and the next layer. Figure 3 represents a 3-layer perceptron having three inputs, two outputs, and the hidden layer including five neurons

The values retrieved from the previous layer are summed up with certain weights, individual for each neuron, plus the bias term [10]. The sum is transformed using the activation function.

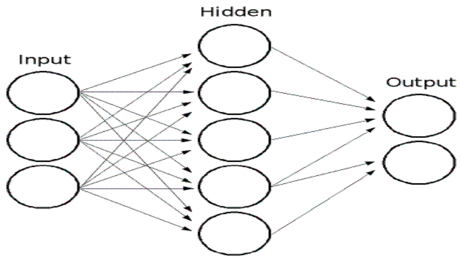

Figure 3: A Multi-Layer perceptron

The perceptron computes a single output from multiple real-valued inputs by forming a linear combination according to its input weights and then putting the output through some nonlinear activation function:
Given output (u_i)

$$u_i = \sum_{j=1}^{n} (w_{i,j}\ x_j + b_i) \tag{4}$$

With the activation function (φ) applied, mathematically the MLP can be written as:

$$y_i = \varphi\left(\sum_{j=1}^{n} (w_{i,j}\ x_j + b_i)\right) \tag{5}$$

where w = weight going to the hidden unit layer
x= Input to hidden unit
b= bias input
φ= Activation function

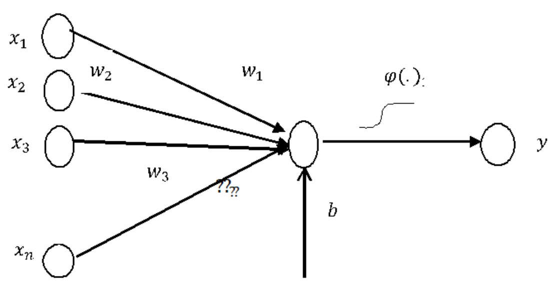

Figure 4. Representation of the MLP equation

A. Learning Algorithm

The MLP uses a backpropagation algorithm to learn and train from the dataset
The back-propagation algorithm is in 2 phases:

- The forward pass phase- computes 'functional signal', feed forward propagation of input pattern signals through network.
- Backward pass phase- computes 'error signal', *propagates* the error *backwards* through network starting at output units (where the error is the difference between actual and desired output values).

Forward pass Algorithm

- Step 1: Initialize weights at random, choose a learning rate η
- Until network is trained:
- For each training example i.e. input pattern and target output(s):
- Step 2: Do forward pass through net (with fixed weights) to produce output(s)
    - i.e., in Forward Direction, layer by layer:
        - Inputs applied
        - Multiplied by weights
        - Summed
        - 'Squashed' by sigmoid activation function
        - Output passed to each neuron in next layer
    - Repeat above until network output(s) produced

Backward pass /Back propagation of error

- Compute error (delta or local gradient) for each
- output unit δ *k*
- Layer-by-layer, compute error (delta or local
- gradient) for each hidden unit δ *j* by backpropagating
- errors (as shown previously)
- Next, update all the weights Δ*wij*
- By gradient descent, and go back to Step 2

The overall MLP learning algorithm, involving forward pass and backpropagation of error (until the network training completion), is known as the Generalized Delta Rule (GDR), or more commonly, the Back Propagation (BP) algorithm

VI. MLP IMPLEMENTATION

The MLP model was implemented on a Personal Computer with 2.30 GHz and 8GB of RAM in Microsoft Windows 10 Operating system platform and Microsoft Excel 2013 with Python Programing Language. The MLP training was defined with parameters epochs = 20, dim_size = 15, num_seq = 30, batch_size = 200, activation function = Sigmoid.

Due to the high imbalance in the datasets, the data were synthetically balanced using the smote method, The datasets 1 and dataset 2 stored in csv format were loaded into python 3.6 IDLE via a read_csv () command. The datasets were divided into two parts (Input and Output). The input data are those with the attributes while the output data contain the target class ('Fraudulent' and 'Normal').

A. Evaluation Metrics

The evaluation of the model was carried out using the various evaluation metrics such as Accuracy, Precision, F1-score, Recall and False alarm rate.

Accuracy: is defined as the number of correct predictions made by the model. It is the proportion of the total number of correct predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

False Alarm Rate (FAR)/False Positive rate: is a ratio of wrongly classified normal instances.

$$False\ Alarm\ Rate = \frac{FP}{TN + FP} \quad (7)$$

Precision: defines the results classified as positive by the model, how many were actually positive. It is the number of items correctly identified as positive out of total true positives.

$$\text{Precision}=\frac{\text{true positive}}{\text{true positive +false positives}} \quad (8)$$

Recall: It is the number of items correctly identified as positive out of the total items classified as positive.

$$\text{Recall}=\frac{\text{true positive}}{\text{true positive+false negatives}} \quad (9)$$

F1-Score: is the weighted average of the precision and the recall, it takes both false negatives and positives into the account and gives a better outlook especially in an uneven class distribution it is given as:

$$\text{F1 Score} = 2(\frac{\text{Precision} * \text{recall}}{\text{Precision} + \text{recall}}) \quad (10)$$

where True positive (TP) represents data detected as fraudulent, True negative (TN) represents data detected as legitimate, False positive (FP) represents normal data detected as fraudulent, and False Negative (FN) is denoted as fraud data detected as normal.

VII. RESULTS

In this section, an evaluation of the study with some metrics is presented with the two datasets. Dataset I reveals the significance of dataset that is characterized with minimum missing data. This is presented in Tables II and IV. The graphical representation of these datasets is presented in Figure 5.

TABLE III: EVALUATION RESULT ON DATASET I

Model	*Accuracy (%)*	*F1 score (%)*	*Precision (%)*	*Recall (%)*	*False Alarm rate (%)*
Multi-Layer Perceptron	96.4	96.3	99.1	93.6	0.001

TABLE IV: EVALUATION RESULT ON DATASET 2

Model	*Accuracy (%)*	*F1 score (%)*	*Precision (%)*	*Recall (%)*	*False Alarm rate (%)*
Multi-Layer Perceptron	77.4	71.4	96.9	56.5	0.002

From the Figure 5 we can conclude that the proposed model performed appreciably better with dataset using the evaluation metrics.

B Performance of Dataset 1 and Datset 2 Using MLP

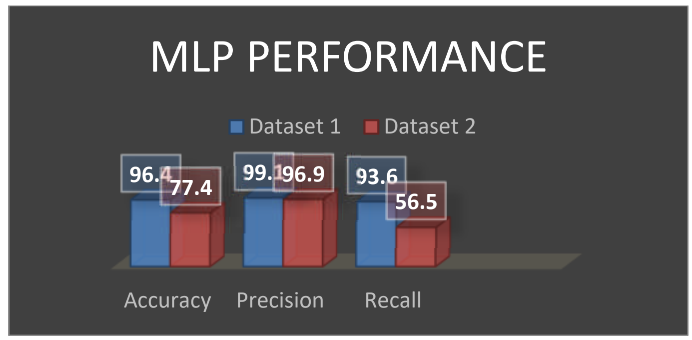

Figure 5.: Performance of Dataset 1 and Dataset 2 using MLP

C Comparative Evaluation

The results of this model were thereafter compared with the results of a work that was implemented using Logistic regression machine learning approach with the same dataset 1 is the result.

TABLE V: ECOMPARATIVE EVALUATION OF MLP AND LOGISTIC REGRESSION

Model	Accuracy (%)	Precision (%)	Recall (%)
Multilayer Perceptron	96.4	99.1	93.6
Logistic Regression	Not given	71	64

This model performed impressively against the performance of the Logistic regression study with the same dataset. Weighted logistic regression was used as an algorithm level technique and random under-sampling was used as data-level technique to build credit card fraud classifier. The classification used in the study was Logistic Regression and the performance metrics are Recall and Precision. A graphical evaluation report of the two models is illustrated in Figure. 6.

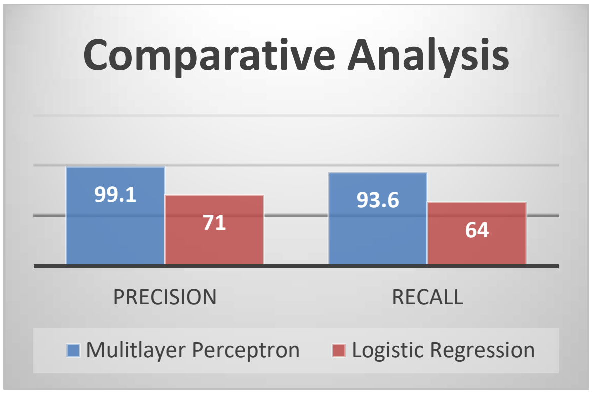

Figure 6: Comparative Analysis of Our Model (MLP) and Logistic Regression

VIII. CONCLUSION

In conclusion, the multilayer perceptron which used information gain method as feature selection technique for obtaining the most relevant features of the dataset was found to be effective in fraud detection; this is hopeful to be of high importance to the financial sector. This study established a fraud detection framework that is capable of unmasking real-time fraudulent transactions. The prediction of the proposed framework records high level of accuracy, precision, recall, good F1-score and very low false alarm rate. In addition, it is observed that the larger dataset, which is Dataset I, yielded high evaluation values than Dataset II, a smaller dataset- Dataset II. This corroborates facts from literatures on the prediction accuracy in big data. Future work will be extended to other algorithms as well as hybridized approach with minimal computational complexity.

REFERENCES

[1] W. S. Albrecht, C. O. Albrecht, C. C. Albrecht and M. F Zimbelman, “Fraud examination,” 5th Edition, Cengage Learning, 2014.

[2] F. N. Ogwueleka, “Data mining application in credit card fraud detection system,” *Journal of Engineering Science and Technology*, 2014, 6(3):311-322.

[3] H. Shao, H. Zhao and G. Chang, “Applying data mining to detect fraud behaviour in customs declaration,” *Proc. of 1st International Conference on Machine Learning and Cybernetics*, 2015, 1241-1244

[4] N. M. Brennan and M. McGrath, “Financial statement fraud: incidents, methods and motives,” *Australian Accounting Review*, 2007, 17(2):49-61.

[5] P. L. Clifton, Vincent, S. Kate and G. Ross, “A comprehensive survey of data mining-based fraud detection research,” School of Business Systems, Faculty of Information Technology, Monash University, Clayton campus, Wellington Road, Clayton, Victoria 3800, Australia, 2012.

[6] C. S. Throckmorton, V. Mohan, J. M. William and C. Leslie, “Financial fraud detection using vocal, linguistic and financial cues,” 2018.

[7] F. Chowdhury and M. S. Ferdous, “Modelling cyber-attacks,” International Journal of Network Security & Its Applications 9(4):13-31, July 2017.

[8] K. Dinaesh, “Cyber defense mathematical modeling and simulation,” *International Journal of Applied Physics and Mathematics, Vol. 2, No. 5, September 2012*

[9] P. Laerte, D. Marcelo, M. Bernardo, F. G. David, and D. T. Deus, “A Formal classification of internet banking attacks and vulnerabilities in combatting financial fraud: A coevolutionary anomaly detection approach,” International Journal of Computer Science & Information Technology (IJCSIT), 2015, Vol 3.

[10] S. Maes, K. Tuyls, B. Vanschoenwinkel, and B. Manderick, “Credit card fraud detection using Bayesian and neural networks,” In *Proceedings of the 1st International naiso congress on neuro fuzzy technologies* 2002, (pp. 261-270).