ID 266

A Machine Learning Approach for Detecting Credit Card Fraudulent Transaction

RMSM Nimashini[1#], RMKT Rathnayake[2] and WU Wickramaarachchi[1]

[1]*Department of Computing, Rajarata University of Sri Lanka, Mihintale, Sri Lanka*
[2]*Department of Physical Sciences and Technologies, Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka*

[#]sachinimilcah@gmail.com

Abstract - The world is reaching a cashless society with the increment of non-cash transactions. E-commerce has become an essential factor in every organization in global trade. Since financial institutions co-operate with billions of online transactions per day, identifying fraudulent transactions has become a challenge. This research was mainly focused on identifying the best intelligent adaptive authentication technique for credit card fraud detection. Areal-world transaction dataset of European credit cardholders and a synthetic dataset were used to extract the historical transactional patterns using Artificial Neural Network (ANN). Different classification algorithms, Logistic Regression, Decision Tree, Random Forest and XGBoost were also used for a comparative analysis to classify a real-world dataset. Among all, ANN and XGBoost have shown the highest performance in the binary classification of fraud and legitimate transactions. ANN has shown an accuracy of 99.94% and high adaptability in handling large datasets, by giving zero misclassification of fraud as a legitimate transaction by reducing the risk to its minimum.

Keywords: ***fraud detection, ANN, adaptive authentication, random forest, decision tree, XGBoost, logistic regression***

I. INTRODUCTION

Financial fraud can be defined as “A deliberate act that is contrary to law, rule, or policy with the intent to obtain an unauthorized financial benefit”. It was reported $24.2 billion was lost worldwide in 2018 due to credit card fraud. As there are millions of credit card users in the world, gross losses from credit card fraud are expected to reach $40 billion in 2027. The Federal Trade Commission (FTC) of America is an organization that protects American consumers and deals with the issues of economic lifestyles. According to the statistics provided by Consumer Sentinel Network Data Book of FTC in 2019, the fraud rate has increased considerably. The FTC has received nearly 271,000 reports from Americans about information misused on an existing account or to open a new credit card account. Figure 1 shows the increment of number of frauds, identity theft, and other fraud reports from 2001 to 2019.

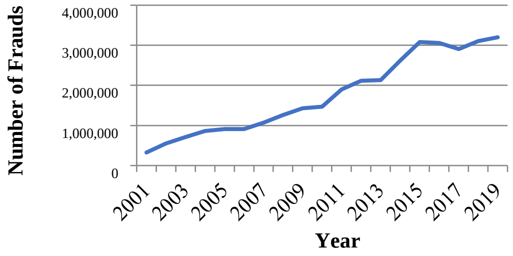


Figure 1 Increment of fraud reports by year Source: FTC, Consumer Sentinel Network Data Book, 2019

According to the categories of identity theft fraud, credit card fraud, loan and lease fraud, phone and utility fraud ranked the top three for several years. Figure 2 shows the statistics of those top three frauds from 2015 to 2019

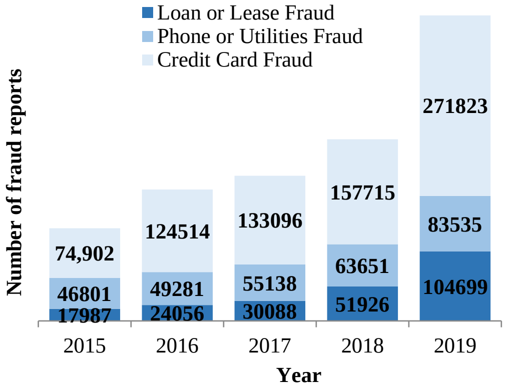


Figure 2. Top three theft reports by year
Source: FTC, Consumer Sentinel Network Data Book, 2019

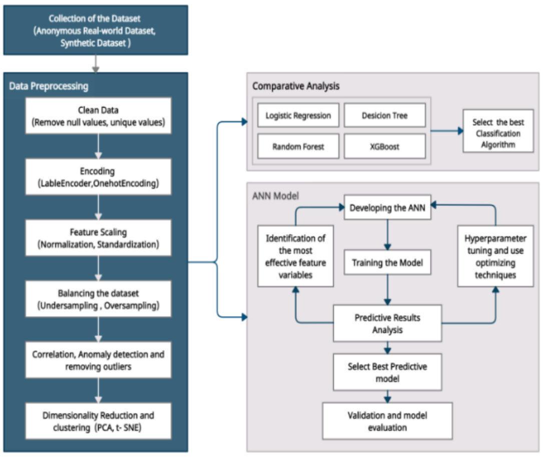
The reported number of Credit Card Fraud has shown a significant increment from 2015 to 2019. Financial institutions have taken many countermeasures to avoid credit card fraud. Different techniques like credit card authorization, Address Verification Systems (AVS), and rule-based fraud detection systems have been used by banking sector for fraud detection. The verification and authentication methods involved in fraud detection cannot identify frauds while they were occurring. The challenge in fraud detection was the dynamic behavior of the fraudsters. Many fraudsters try to behave like legitimate users. So the predictive systems should be constantly updated with the transaction behaviour.

By conducting this research, it is expected to identify credit card frauds by considering large historical data of the user's transaction behavior. The system should be intelligent to identify highly changing fraud styles using data mining and machine learning techniques with the help of historical transaction datasets.

Different machine learning approaches and classification algorithms were used by many researchers for credit card fraud detection based on probability. Naïve Bayes classifier, K-Nearest Neighbor (KNN), Fuzzy Logic, Bayesian Network, KNN, SVM, Decision Tree, Hidden Markov Model (HMM), and Logistic Regression was commonly used in fraud detection. Many classification algorithms are not much capable of identifying novel transaction patterns. Also, found that they are not capable to process or not scalable to large datasets when compared to neural networks. Many researchers have focused on developing credit card fraud detection systems using neural networks. It was identified that ANN has shown better results in credit card fraud detection and they are highly adaptive and perform well in detecting novel credit card frauds.

II. METHODOLOGY AND EXPERIMENTAL DESIGN

The methodology used in this research mainly consists of 3 parts. Data acquisition and pre-processing, Comparative analysis of classification algorithms, and the ANN model development. Different pre-processing techniques like data cleaning, encoding, feature scaling, data balancing, correlation, outlier removal, dimensionality reduction and clustering were used. Exploratory Data Analysis was used to identify the distribution and relationships of data. The main part of the methodology is the ANN model building for prediction. The other part is the comparative analysis of different classification algorithms. Logistic Regression, Decision Tree, Random Forest, and XGBoost were used and analyzed to identify the best classification technique. Figure 3 has shown the flow of used methodology.



Collection of the Dataset (Anonymous Real-world Dataset, Synthetic Dataset)
Data Preprocessing
Clean Data (Remove null values, unique values)
Encoding (LableEncoder,OnehotEncoding)
Feature Scaling (Normalization, Standardization)
Balancing the dataset (Undersampling , Oversampling)
Correlation, Anomaly detection and removing outliers
Dimensionality Reduction and clustering (PCA, t- SNE)
Comparative Analysis
Logistic Regression
Desicion Tree
Random Forest
XGBoost
Select the best Classification Algorithm
ANN Model
Developing the ANN
Identification of the most effective feature variables
Training the Model
Hyperparameter tuning and use optimizing techniques
Predictive Results Analysis
Select Best Predictive model
Validation and model evaluation

Figure 3. Methodology

A. Dataset

In this research, mainly two datasets were used, a real-world dataset (Dataset 1) and a synthetic dataset (Dataset 2). The dataset (1) has real-world transactions that were previously transformed into Principal Component Analysis (PCA) due to confidentiality. Those transactions were done by European Credit Cardholders in 2013 within 2 days. It has 31 attributes where 28 attributes are anonymous. The dataset contains a total number of 284,807 transactions. The number of legitimate transactions was 284,315 which was 99.83% and the number of fraudulent transactions was 492 which was 0.172% from the whole dataset. It consists of 31 features and only 3 features were disclosed. The following Table 1 gives the information on the dataset (1).

Table 1. Details on the attributes of dataset(1)

No.	Attributes	Description
1	Time	The seconds elapsed between each transaction
2	Amount	Transaction Amount
3	Class	Target Variable (Fraud or legitimate)
4	Other 28 features. (V1, V2,..V28)	Anonymous variables which were transformed into PCA

Dataset (2) is a synthetic dataset that is artificially generated and created algorithmically for research purposes. It consists of 2627 legitimate transactions and 448 fraudulent transactions. This dataset has 11 attributes. Following Table 2 gives details about attributes in the dataset and their description. Dataset (2) is a synthetic dataset that is artificially generated and created algorithmically for research purposes. It consists of 2627 legitimate transactions and 448 fraudulent transactions. This dataset has 11 attributes. Following Table 2 gives details about attributes in the dataset and their description.

Table 2. Details on the attributes of dataset (2)

No.	Attributes	Description
1	Merchant_id	Unique identity.
2	Average_amount	The average transaction amount.
3	Transaction_amount	The transaction amount
4	Is_declined	Whether the transaction is previously declined or not (yes/no)
5	TotalNumberof declines_day	Total number of declined happened within a day.
6	isForeignTransaction	Whether the transaction is a foreign transaction or not (yes/no)
7	isHighRiskCountry	Whether the transaction is started from a high risk country (yes/no)
8	Daily_chargeback_avg_amt	Average chargeback amount per day.
9	6_month_avg_chbk_amt	Average chargeback amount per 6 months.
10	6_month_chbk_freq	Chargeback frequency within 6 months.
11	isFradulent	Target Variable

B. *Data Preprocessing*

Different feature engineering techniques were used to preprocess the datasets. Handling of missing values, encoding strings into numerical, applying scaling techniques like normalization and standardization, and data balancing was mainly used. Both datasets were highly imbalanced as the number of fraudulent transactions was very low when compared to the number of legitimate transactions. To reduce the skewness of the model towards the highest data population the dataset should be balanced using sampling techniques. The Undersampling techniques are not suitable as they reduce the sample size of the dataset. Therefore Oversampling techniques were used to synthesize the number of fraudulent transactions. The highest accuracy was reached when the imbalanced nature of the dataset was handled by using the Synthetic Minority Oversampling Technique (SMOTE).

C. *Exploratory Data Analysis(EDA)*

The balanced datasets were further processed to identify the relationships among feature variables, to visualize the distribution of variables, to identify correlations, and to identify clusters. The Pearson correlation distribution was used to measure the statistical relationship, or association, between two continuous variables. Positive and negative correlation of a feature with the target class was used to learn which features heavily influence the identification of a specific transaction as a fraud. The following Figure 4 shows the Pearson correlation heatmap obtained for the dataset (1).

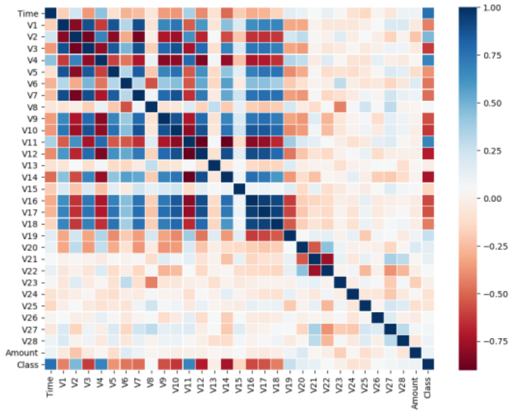


Figure 4. Correlation heatmap for dataset (1)

According to the correlation heatmap, the variables V2, V4, V11, and V19 have shown strong positive correlations with the target class. This means that the higher the value for one of these features, the more likely it will be a fraud transaction. Features V10, V12, V14, and V16 have shown strong negative correlations with the target class. This means that the lower the value for one of these features, the more likely it will be a fraud transaction. The following Figure 5 shows the Pearson correlation heatmap obtained for the dataset (2).

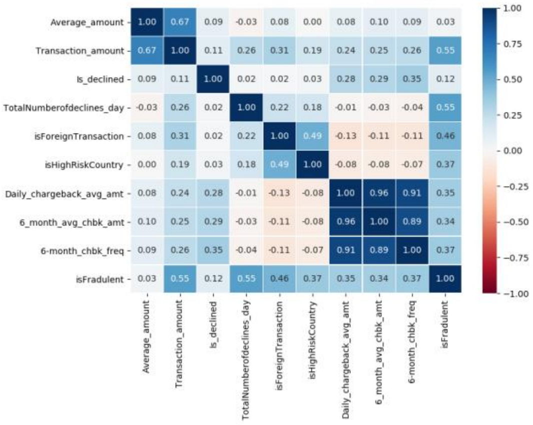


Figure 5. Correlation heatmap for dataset (2)

D. Detection and Treating Outliers

As these highly correlated variables have a high impact on the prediction of the target class, the extreme outliers in these selected variables should be identified and removed to improve the accuracy of the model. The presence of outliers can be determined by observing the distribution of selected feature variables which are positively and negatively correlated with the target class of both datasets. The following Figure 6 shows the distribution of positively correlated features of the dataset (1).

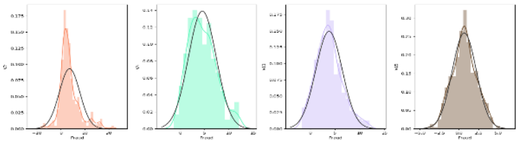


Figure 6. Distribution of positively correlated features (Dataset 1)

According to the above distributions, the V2 variable was having data points with a huge difference from the normal distribution of other data points. The V4, V11, and V19 variables show few data points which are deviated from the normal distribution. To observe the extreme outliers of the variables boxplot diagrams can be used. Figure 7 shows the boxplots for the visualization of present outliers of dataset (1).

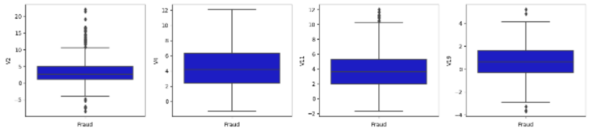


Figure **7**. Boxplot of positively correlated features (Dataset 1)

The above boxplots of variables V2, V4, V11 and V19 shows some outliers of faruds positioning more above third quartile and more below first quartile. These extreme outliers were removed from these variables as they can affect the results in machine learning. The following Figures 8 and 9 show the distribution of negatively correlated features and the boxplots of the dataset (1).

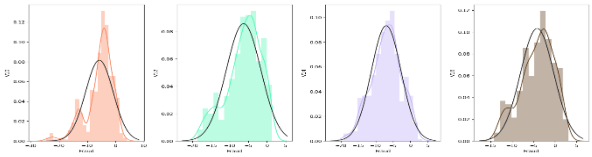


Figure 8. Distribution of negatively correlated features with fraud class (Dataset 1)

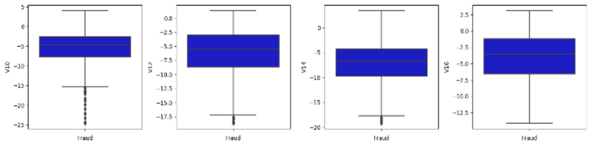


Figure 9. Boxplots of negatively correlated features with fraud class (Dataset 1)

Among the above distribution of V10, V12, V14, and V16 variables, V10 shows many data points which are deviated from the normal distribution. Those outlilers may effect in misclassification and that should be removed. The following Figure 10 shows the distribution of the features which have a strong positive correlation with the target class of dataset (2) and Figure 11 shows their boxplots representation.

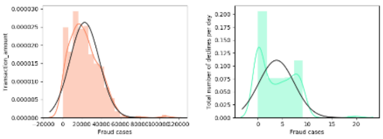


Figure 10. Distribution of positively correlated features (Dataset 2)

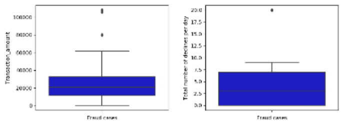


Figure 11. Distribution of positively correlated features (Dataset 2)

According to the above distribution both variables have extreme outliers and those should be removed before using them in machine learning model.

For removal of outliers, Interquartile Range Method (IQR) was used. It measures the variability by dividing the dataset into quartiles. The quartiles were identified by dividing the dataset into 4 equal parts after sorting into ascending order. The four quartiles Q1, Q2, and Q3 represent the 25th percentile, 50th percentile, and 75th percentile of the data respectively. The IQR is calculated using the difference between the 75th and the 25th percentiles of the data (IQR = Q3 – Q1). The IQR was used to identify outliers by defining limits on the sample values that are a factor k of the IQR below the 25th percentile or above the 75th percentile. The common value for the factor k = 1.5 was used for the calculation. The data points which are below Q1 – 1.5*IQR or above Q3 + 1.5 IQR were considered as the outliers of the dataset. The identified outliers for the selected features were removed from the dataset to increase the quality of the dataset. Using the IQR method total number of 121 outliers from dataset (1) and 6 extreme outliers from dataset (2) were removed.

E. Cluster Identification Using Dimensionality

Reduction t-distributed stochastic neighbor embedding (t-SNE) was used to identify the clusters of the dataset by dimensionality reduction. t-SNE measures the euclidean distance between two points and then plots that distance on a normal curve that is centered on the point of interest. Lastly, it takes the distance between point 2 and where it is on the normal curve. Figure 13 shows the t-SNE distribution of the dataset (1) and Figure 14 shows the t-SNE distribution of the dataset (2) respectively.

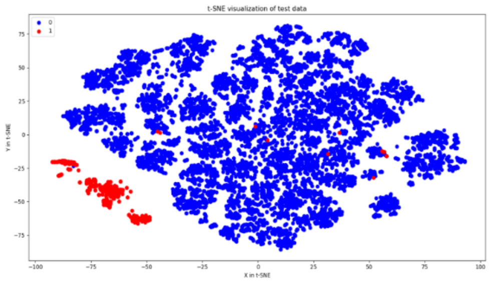


Figure 15. t-SNE distribution of dataset (1)

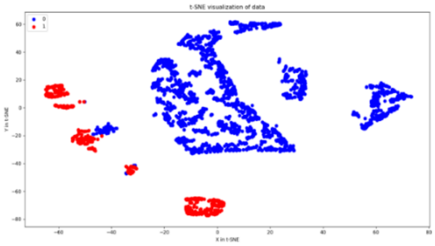


Figure 14. t-SNE distribution of dataset (2)

According to the distribution of data points, and the available clusters, it can be observed that the target classes are clearly separable in both datasets. Therefore it can be used for further processing with machine learning models and classification algorithms.

F. Artificial Neural Network (ANN)

An ANN was used for the prediction of transactions by extracting the hidden patterns. The Keras sequential model was used to create the network architecture. The Dense class was defined to create a fully connected network structure with layers. The input layer has the exact number of input features. The input dimension was set according to the number of feature variables that were ready to feed into the neural network. In this research, the input dimensions were set as 29 and 8 for dataset (1) and dataset (2) respectively. Each layer has a specific number of neurons and a defined activation function. The Rectified Linear Unit Activation function was used (ReLU) on the input layer and hidden layers. The sigmoid activation function was used to ensure the network output is between 0 and 1. The classification was done using a default threshold of 0.5. The loss function was used to evaluate the set of weights and the optimizer was used to search through different weights. The cross-entropy was used as the loss argument defined in Keras as "binary_crossentropy". The optimizer was defined as the efficient stochastic gradient descent algorithm "adam".The model evaluation was done using evaluate() function. It returns the loss and the accuracy of the model on the dataset. The predict() function was used to get prediction probability in the range between 0 and 1 as the sigmoid function gives in the output layer.

III. RESULTS

A. *Predictive Analysis of ANN*

The ANN has generated the results for dataset (1) with an accuracy of 99.94% and loss of 0.0032. The confusion matrix for the balanced dataset has given the classification showing an 82877 of True Positive, 103 of False Positive, 83285 True Negative, and 0 False Negative cases. For the whole dataset, the confusion matrix has given 276862 of True Positive, 246 of False Positive, 371 of True Negative, and 0 False Negative Cases. Classification of fraud transactions with zero false negative cases has reduced the risk of identifying a fraudulent transaction as a legitimate transaction with 100% accuracy. Identification of a legitimate transaction as fraudulent is also can be used in further processing as their risk of being a fraudulent transaction is high to a certain extent. That can be added to a flagged fraud list to use in future predictions. Following Figure 15 represent the learning curves of accuracy and loss graph obtained from the ANN of the dataset (1).

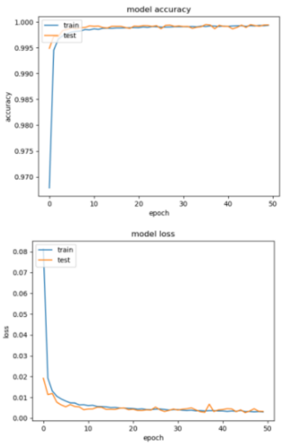


Figure 15. Model accuracy and loss curve for dataset (1)

For the ANN of dataset (2), the balanced dataset has given the classification of binary classes showing a 758 of True Positive, 32 of False Positive, 778 True Negative, and 9 False Negative cases. For the whole dataset, it has given 2520 of True Positive, 107 of False Positive, 435 of True Negative and 7 False Negative Cases. When the batch size and epoch combination were 10 and 100 it has given an accuracy of 96.83% and a loss of 0.114. When increasing the batch size from 25 to 32 and keeping the range of epochs from 50 – 100 , the highest accuracy was obtained when the batch size was 30 and when the number of epochs was 75. The accuracy was reached up to 97.40% and able to reduce the loss up to 0.07485. Following Figure 16 represent the learning curves of accuracy and loss graph obtained from the ANN of the dataset (2).

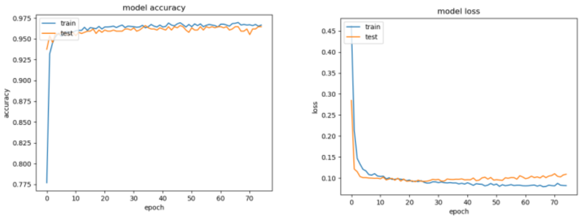


Figure 16. Model accuracy and loss curve for dataset (2)

For comparative analysis mainly four classification algorithms, Logistic Regression, Decision Tree, Random Forest, and XGBoost were used with dataset (1). The following Table 3 gives the information of the classification report that was obtained when using each algorithm.

Table 3. Classification report

	Logistic Regresion	Decisin Tree	Randm Forest	XG Boost
Accuracy	**97.16 %**	**96.35 %**	**96.43 %**	**99.94 %**
Precision 0	0.96	0.95	0.94	1.00
1	0.98	0.98	0.99	1.00
Recall 0	0.98	0.98	0.99	1.00
1	0.96	0.95	0.93	1.00
F1-score 0	0.97	0.96	0.97	1.00
1	0.97	0.96	0.96	1.00

For Logistic Regression 52 of false negatives, for Decision Tree 50 of false negatives, for Random Forest 62 of false negatives, and for XGBoost only 6 of false negatives was recorded. The XGBoost has shown the minimum number of false negatives with the highest accuracy of 99.94% and it can be identified as the best classification algorithm for credit card fraud detection.

IV. DISCUSSION AND CONCLUSION

According to the obtained results, adaptive authentication using ANN has shown a 99.94% accuracy with zero false negative cases for dataset (1). It can be concluded that the risk of misclassification of fraud transactions as legitimate has reduced by 100%. The false positive cases can be used in future prediction as they can be categorized into flagged fraud transactions. The ANN for dataset (2) was able to 97.40% with only 7 misclassified false negative cases. The Logistic Regression has shown 97.16% accuracy, Decision Tree has shown 96.35% accuracy, Random Forest has shown 96.43% accuracy and XGBoost has shown 99.94% accuracy, which is equal to the accuracy of ANN. But the value in ANN was increasing with the use of huge datasets, as the classification algorithms are not capable of handling and adapting to the huge datasets.

Using obtained results of the research and comparative analysis of classification algorithms, it can be concluded that the developed ANN has the highest capability in providing adaptive authentication for credit card fraud detection with the highest accuracy of 99.94%. As future suggestions, this model can be improved and used if the real world dataset is more disclosed to easy access and if it is not anonymous.

REFERENCES

Eng PJM, (2020) Credit Card Fraud Detection using PSO Optimized Neural Network, International Journal of Engineering and Advanced Technology, 9, 360-363.

Fahmi M, Hamdy A, Nagati K, (2016) Data Mining Techniques for Credit Card Fraud Detection : Empirical Study, Sustainable Vital Technologies in Engineering & Informatics, 1- 9.

Federal Trade Commission 2019, "Consumer Sentinel Network DataBook",2019.[Online].Available:https://www.ftc.gov/reports/consumer-sentinel-network-data-book-2019

Maniraj SP, Saini A, Ahmed S, Sarkar SD, (2019) Credit Card Fraud Detection using Machine Learning and Data Science, International Journal of Engineering Research & Technology (IJERT) , 8,110-115.

Mishra C, Gupta DBL, Singh R, (2017) Credit Card Fraud Identification Using Artificial Neural Networks, International Journal of Computer Systems, ISSN, 4, 2394-1065.

Nadim A, Sayem I, (2019) Mutsuddy A, Chowdhury MS, Analysis of machine learning techniques for credit card fraud detection, in Proc. of Intern. Conf. on Machine Learning and Data Engineering, ICMLDE, 42-47.

Pozzolo AD, Boracchi G, Caelen O, Alippi C, Bontempi G, (2018) Credit card fraud detection: A realistic modeling and a novel learning strategy, IEEE Transactions on Neural Networks and Learning Systems, 29, 3784-3797.

Priscilla C, Prabha D, (2019) Credit Card Fraud Detection: A Systematic Review, 402-407.

Sorournejad S, Zojaji Z, Atani R, Monadjemi A.H, (2016) A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective.

Zareapoor M, Shamsolmoali P, (2015) Application of credit card fraud detection: Based on bagging ensemble classifier, Procedia Computer Science, 48,679-685.

Zhou X, Zhang Z, Wang L, Wang P, (2019) A Model Based on Convolutional Neural Network for Online Transaction Fraud Detection, in Proc. of the Intern. Joint Conference on Neural Networks .

AUTHOR BIOGRAPHIES

R.M.S.M.Nimashini is a fourth year undergraduate of BSc. (Hons) degree in Information Technology at Rajarata University of Sri Lanka. Currently following an internship in Analytics and Data science in London Stock Exchange Group (LSEG) in Sri Lanka. My research interests are machine learning, business intelligence, and data mining.

Prof. (Dr.) R.M.K.T. Rathnayake is a professor in Statistics at Sabaragamuwa university of Sri Lanka. His research interest are financial mathematics, Time Series Modelling and Data Mining and Machine Learning, Big Data Analytics, Business Models, Machine learning and Multi objective Combinatorial optimization.

Wiraj Udara Wikramaarachchi is a lecturer at Rajarata University of Sri Lanka. Currently reading D.Eng. in Computer Science and Technology – Wuhan University of Technology, in China. His research interests are Biometrics, Information Security, Privacy Protection, Image processing and Natural Language Processing.

.