# Credit Card Fraud Detection Using Artificial Neural Network

**Sachin S**
Dept of M.Sc
Dr.M.G.R Educational and Research Institute

***Abstract-*** *It is vital that credit card companies are able to identify fraudulent credit card transactions so that customers are not charged for items that they did not purchase. Such problems can be tackled with Data Science and its importance, along with Machine Learning, cannot be overstated. This project intends to illustrate the modeling of a data set using machine learning with Credit Card Fraud Detection. The Credit Card Fraud Detection Problem includes modelling past credit card transactions with the data of the ones that turned out to be fraud. This model is then used to recognize whether a new transaction is fraudulent or not. The objective here is to detect 100% of the fraudulent transactions while minimizing the incorrect fraud classifications. In the classification process, focused on analyzing and pre-processing data sets as well as the deployment of multiple anomaly detection algorithms such as Local Outlier Factor and Isolation Forest algorithm on the Credit Card Transaction data. So will make use of accuracy and precision to evaluate the performance of the proposed system.*

***Keywords***- Credit card fraud, applications of machine learning, data science, isolation forest algorithm, local outlier factor, automated fraud detection.

## I. INTRODUCTION

Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people. Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve. Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs. Any technology user today has benefitted from machine learning. Facial recognition technology allows social media platforms to help users tag and share photos of friends. Optical character recognition (OCR) technology converts images of text into movable type. Recommendation engines, powered by machine learning, suggest what movies or television shows to watch next based on user preferences. Self-driving cars that rely on machine learning to navigate may soon be available to consumers. Machine learning is a continuously developing field. Because of this, there are some considerations to keep in mind as you work with machine learning methodologies, or analyze the impact of machine learning processes.

Illegal use of credit card or its information without the knowledge of the owner is referred to as credit card fraud. Different credit card fraud tricks belong mainly to two groups of application and behavioral fraud. Application fraud takes place when, fraudsters apply new cards from bank or issuing companies using false or other's information. Multiple applications may be submitted by one user with one set of user details (called duplication fraud) or different user with identical details (called identity fraud). Behavioral fraud, on the other hand, has four principal types: stolen/lost card, mail theft, counterfeit card and „card holder not present" fraud. Stolen/lost card fraud occurs when fraudsters steala credit card or get access to a lost card. Mail theft fraud occurs when the fraudster get a credit card in mail or personal information from bank before reaching to actual cardholder. In both counterfeit and „card holder not present" frauds, credit card details are obtained without the knowledge of card holders. In the former, remote transactions can be conducted using card details through mail, phone, or the Internet. In the latter, counterfeit cards are made based on card information.

'Fraud' in credit card transactions is unauthorized and unwanted usage of an account by someone other than the owner of that account. Necessary prevention measures can be taken to stop this abuse and the behaviour of such fraudulent practices can be studied to minimize it and protect against similar occurrences in the future. In other words, Credit Card Fraud can be defined as a case where a person uses someone else's credit card for personal reasons while the owner and the card issuing authorities are unaware of the fact that the card is being used.

Fraud detection involves monitoring the activities of populations of users in order to estimate, perceive or avoid objectionable behaviour, which consist of fraud, intrusion, and defaulting. This is a very relevant problem that demands the attention of communities such as machine learning and data science where the solution to this problem can be automated. This problem is particularly challenging from the perspective of learning, as it is characterized by various factors such as class imbalance. The number of valid transactions far outnumber fraudulent ones.

Also, the transaction patterns often change their statistical properties over the course of time These are not the only challenges in the implementation of a real-world fraud detection system, however. In real world examples, the massive stream of payment requests is quickly scanned by automatic tools that determine which transactions to authorize. Machine learning algorithms are employed to analyse all the authorized transactions and report the suspicious ones. These reports are investigated by professionals who contact the cardholders to confirm if the transaction was genuine or fraudulent. The investigators provide a feedback to the automated system which is used to train and update the algorithm to eventually improve the fraud-detection performance over time.

Outlier detection is the identification of objects, events or observations which do not conform to an expected pattern or other items in a dataset. As one of the important tasks of data mining, outlier detection is widely used in the fields of network intrusion detection, medical diagnosis, industrial system fault, flood prediction and intelligent transportation system. Many existing research methods about outlier detection are divided into the following categories: distribution-based methods, distance-based methods, density-based methods, and clustering methods. Specifically, the distribution-based method needs to obtain the distribution model of data to be tested in advance, which depends on the global distribution of the dataset, and is not applicable to the dataset with uneven distribution. The distance-based approach requires users to select reasonable distance, scale parameters and is less efficient on high-dimensional datasets.

In the clustering method, the outlier is not the target of the cluster resulting that the abnormal point cannot be accurately analyzed. The above outlier detection methods all adopt global anomaly standards to process data objects, which cannot perform on the datasets with uneven distribution. In practical applications, the distribution of data tends to be skewed, and there is a lack of indicators that can classify data. Even if tagged datasets are available, their applicability to outlier detection tasks is often unknown. The density-based local outlier detection method can effectively solve the above problems by describing the degree of outliers of data points quantified by local density.

## II. LITERATURE SURVEY

Due to the rapid growth in e-business and electronic payment systems, Fraud is rising in banking transactions associated with credit cards. This paper intends to develop credit card fraud detection (CCFD) model based on Artificial Neural Networks (ANN) and Meta Cost procedure to reduce risk reputation and risk of loss. ANN strategy have been used for credit card fraud prevention and detection. Because of the unbalanced nature of the data (Fraud and Non-Fraud cases), the detection of fraudulent transactions is difficult to achieve. To deal with the problem of imbalanced data, Meta Cost procedure is added. The proposed model, which is called Cost Sensitive Neural Network (CSNN), is based on misuse detection approach. Compared to the model based on Artificial Immune System (AIS), this model showed cost saving and increased detection rate. Data of this study is taken from real transactional data provided by a big Brazilian credit card issuer.

Credit card fraud is a serious problem in financial services. Billions of dollars are lost due to credit card fraud every year. There is a lack of research studies on analyzing real-world credit card data owing to confidentiality issues. In this paper, machine learning algorithms are used to detect credit card fraud. Standard models are first used. Then, hybrid methods which use AdaBoost and majority voting methods are applied. To evaluate the model efficacy, a publicly available credit card data set is used. Then, a real-world credit card data set from a financial institution is analyzed. In addition, noise is added to the data samples to further assess the robustness of the algorithms. The experimental results positively indicate that the majority voting method achieves good accuracy rates in detecting fraud cases in credit cards.

Fraud can be defined as wrongful or criminal deception intended to result in financial or personal gain, or to damage another individual without necessarily leading to direct legal consequences. The two main mechanisms to avoid frauds and losses due to fraudulent activities are fraud prevention and fraud detection systems. Fraud prevention is the proactive mechanism with the goal of disabling the occurrence of fraud. Fraud detection systems come into play when the fraudsters surpass the fraud prevention systems and start a fraudulent transaction. With the developments in information technology and improvements in communication channels, fraud is spreading all over the world, resulting in huge financial losses. Though fraud prevention mechanisms

such as CHIP&PIN are developed, these mechanisms do not prevent the most common fraud types such as fraudulent credit card usages over virtual POS terminals through Internet or mail orders. As a result, fraud detection is the essential tool and probably the best way to stop such fraud types. In this study, classification models based on Artificial Neural Networks (ANN) and Logistic Regression (LR) are developed and applied on credit card fraud detection problem. This study is one of the firsts to compare the performance of ANN and LR methods in credit card fraud detection with a real data set.
It is easy enough to be infected with communicable and vector-borne diseases, which have very similar symptoms, most of which occur after days. Nowadays technology can help in the correct diagnosis of these diseases. Early diagnosis is necessary to ensure that appropriate treatments and medications are administered, which requires the need for an automated system to predict possible infections. This requires a system that allows the patient to distinguish between these conditions and diagnose the possible disease based on symptoms. After having diagnosed the disease, the goal is to provide appropriate treatment based on the type of disease expected. The implementation of this medical diagnosis system is carried out with the help of Artificial Neural Networks that use backpropagation algorithm for training. With the implementation of Artificial Neural Networks in medical diagnosis, the accuracy of the system improves with respect to the rule-based model and with the use of the backpropagation algorithm together with the gradient optimization technique, the results are more precise.

## III. PROPOSED SYSTEM

The kaggle datasets are trained by using the SMOTE technique. SMOTE technique is used to solve data imbalance problem. Using the smote technique the data, which is nothing but the transactions are trained. This technique is mainly used to differentiate the fraud transactions from the original transactions done by the card holders. Finally, the smote provide the balance data.

To improve the accuracy level of the balance data uses the latest machine learning algorithms to detect anomalous activities, called outliers. Local Outlier Factor It is an Unsupervised Outlier Detection algorithm. 'Local Outlier Factor' refers to the anomaly score of each sample. It measures the local deviation of the sample data with respect to its neighbours. More precisely, locality is given by k-nearest neighbours, whose distance is used to estimate the local data. The Isolation Forest 'isolates' observations by arbitrarily selecting a feature and then randomly selecting a split value between the maximum and minimum values of the designated feature. Recursive partitioning can be represented by a tree, the number of splits required to isolate a sample is equivalent to the path length root node to terminating node. The average of this path length gives a measure of normality and the decision function which use.

First of all, we obtained our dataset from Kaggle, a data analysis website which provides datasets. Inside this dataset, there are 31 columns out of which 28 are named as v1-v28 to protect sensitive data. The other columns represent Time, Amount and Class. Time shows the time gap between the first transaction and the following one. Amount is the amount of money transacted. Class 0 represents a valid transaction and 1 represents a fraudulent one.

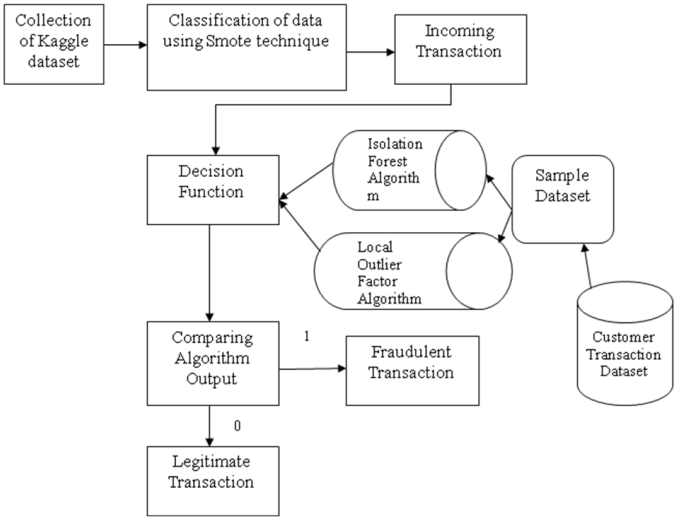


Fig 1.1 System Architecture

### SYSTEM MODULES

### CREDIT CARD DATA

The kaggle is an online community that allows the user to find and publish the datasets. The datasets used in the CCFD system contains transactions made by credit cards by credit card holders. First of all, we obtained our dataset from Kaggle, a data analysis website which provides datasets. Inside this dataset, there are 31 columns out of which 28 are named as v1-v28 to protect sensitive data.

### SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE

SMOTE (Synthetic minority oversampling technique) is a machine learning technique used for classification of data. The kaggle datasets are trained by using the SMOTE technique. SMOTE technique is used to solve data imbalance problem. Using the smote technique the data, which is nothing but the transactions are trained. This

technique is mainly used to differentiate the fraud transactions from the original transactions done by the card holders. Initially the transaction data are stored in a confluence form. Thus the confluence data have been trained by the SMOTE technique to synthesize the fraud transactions from the non fraud transactions. The synthetic minority oversampling technique shrinks the fraud transaction from the non-fraud transactions.

**LOCAL OUTLIER FACTOR**

It is an Unsupervised Outlier Detection algorithm. 'Local Outlier Factor' refers to the anomaly score of each sample. It measures the local deviation of the sample data with respect to its neighbours. More precisely, locality is given by k-nearest neighbours, whose distance is used to estimate the local data.

**ISOLATION FOREST ALGORITHM**

The Isolation Forest 'isolates' observations by arbitrarily selecting a feature and then randomly selecting a split value between the maximum and minimum values of the designated feature. Recursive partitioning can be represented by a tree, the number of splits required to isolate a sample is equivalent to the path length root node to terminating node. The average of this path length gives a measure of normality and the decision function which use.

**ALGORITHM**

**LOCAL OUTLIER FACTOR (LOF)**

The LOF algorithm is defined iby using density-based methods. For each data point, the process of finding the LOF includes calculating the degree of outlying. The idea of a local outlier is introduced by the LOF. The key definitions for the LOF are:

**k-distance of a data point p.**

The distance between the two data points p and o can be calculated by using a Euclidean n-dimensional space.

**Reachability Density (Rd)**

It is defined as the maximum of K-distance of Xj and the distance between Xi and Xj. The distance measure is problem-specific (Euclidean, Manhattan, etc.).

**Local Reachability Density (Lrd)**

LRD is inverse of the average reachability distance of A from its neighbors. Intuitively according to LRD formula, more the average reachability distance (i.e., neighbors are far from the point), less density of points are present around a particular point. This tells how far a point is from the nearest cluster of points. Low values of LRD implies that the closest cluster is far from the point.

**Local Outlier Factor (Lof)**

LRD of each point is used to compare with the average LRD of its K neighbors. LOF is the ratio of the average LRD of the K neighbors of A to the LRD of A.

Intuitively, if the point is not an outlier (inlier), the ratio of average LRD of neighbors is approximately equal to the LRD of a point (because the density of a point and its neighbors are roughly equal). In that case, LOF is nearly equal to 1. On the other hand, if the point is an outlier, the LRD of a point is less than the average LRD of neighbors. Then LOF value will be high.

Generally, if LOF> 1, it is considered as an outlier, but that is not always true. Let's say we know that we only have one outlier in the data, then we take the maximum LOF value among all the LOF values, and the point corresponding to the maximum LOF value will be considered as an outlier.

**ISOLATION FOREST**

Isolation forest detects anomalies by randomly partitioning the domain space. Yeah, you're heard me right- It works similar to Decision trees algorithm, where we start with a root node and keep on partitioning the space. In Isolation forest we partition randomly, unlike Decision trees where the partition is based on Information gain. Partitions are created by randomly selecting a feature and then randomly creating a split value between the maximum and the minimum value of the feature. Isolation forest is an ensemble method. So we create multiple Isolation trees(generally 100 trees will suffice) and we take the average of all the path lengths. This average path length will then decide whether a point is anomalous or not.

**Anomaly score**

Anomaly score is given by the following formula

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}}$$

$$c(n) = 2H(n-1) - ({}^{2(n-1)}/_{n})$$

Where
n- Number of data points
c(n)- It is the average path length of unsuccessful search in a Binary search tree.

## IV. SCREEN SHOTS

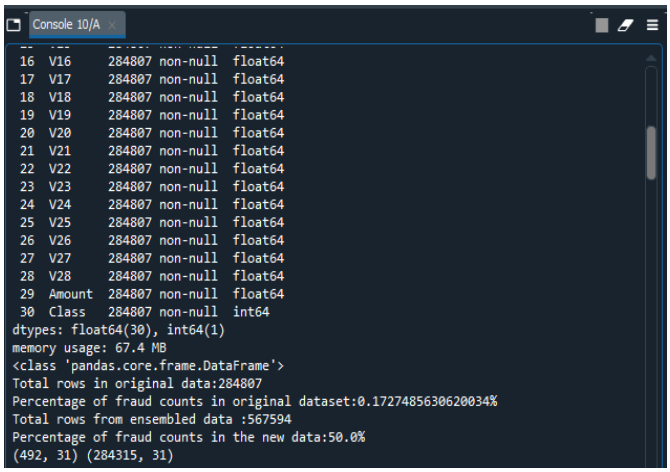


Fig 1.2 Data Pre-processing

The datasets contains transactions made by credit cards by cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions. It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2,…, V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-senstive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise. I use the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE balances the class distribution by creating new synthetic instances of the minority class.

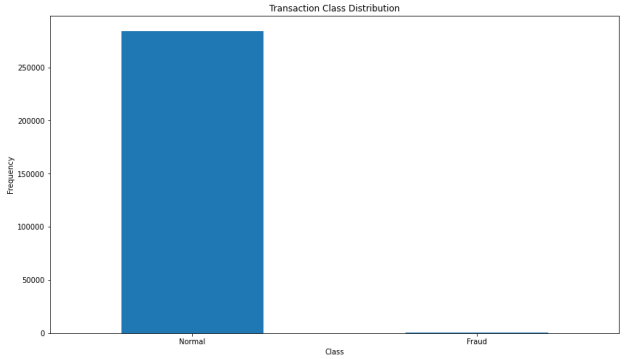


Fig 1.3 Transaction Class Distribution

This graph shows that the number of fraudulent transactions is much lower than the legitimate ones.

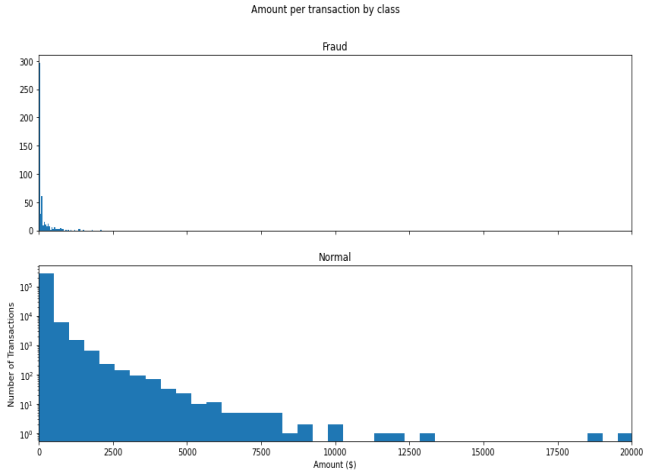


Fig 1.4 Amount per transaction by Class

This graph represents the amount that was transacted. A majority of transactions are relatively small and only a handful of them come close to the maximum transacted amount.

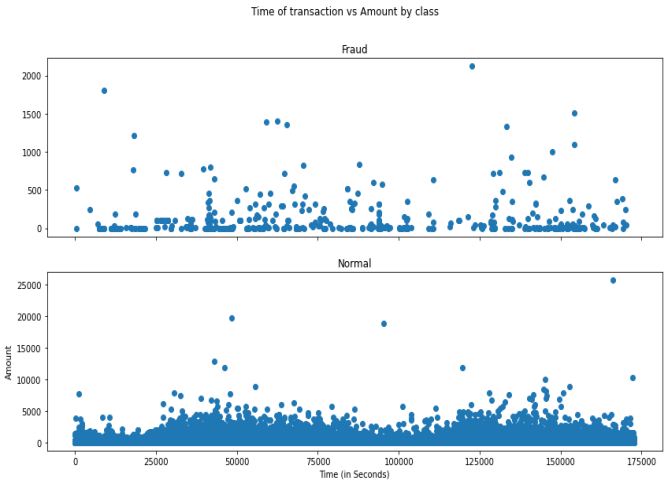


Fig 1.5 Time of transaction vs amount by class

This graph shows the times at which transactions were done within two days. It can be seen that the least

number of transactions were made during night time and highest during the days.


Fig 1.6 Classification Result

The above figure shows the classification result of creditcard fraud detection. The fraud cases is 49 and valid cases are 28432.

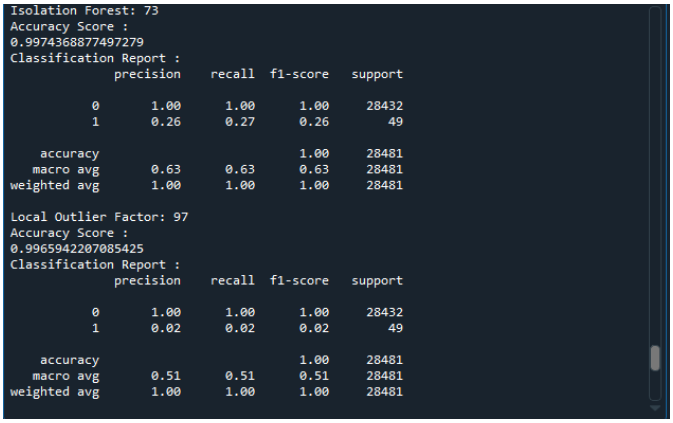


Fig 1.7 Performance of Classification Algorithms

Isolation Forest detected 73 errors versus Local Outlier Factor detecting 97 errors. Isolation Forest has a 99.74% more accurate than LOF of 99.65%. When comparing error precision & recall for 3 models , the Isolation Forest performed much better than the LOF as we can see that the detection of fraud cases is around 27 % versus LOF detection rate of just 2 %. So overall Isolation Forest Method performed much better in determining the fraud cases which is around 30%.

## V. CONCLUSION

Credit card fraud is without a doubt an act of criminal dishonesty. This article has listed out the most common methods of fraud along with their detection methods and reviewed recent findings in this field. This paper has also explained in detail, how machine learning can be applied to get better results in fraud detection along with the algorithm, pseudocode, explanation its implementation and experimentation results. While the algorithm does reach over 99.6% accuracy, its precision remains only at 28% when a tenth of the data set is taken into consideration. However, when the entire dataset is fed into the algorithm, the precision rises to 33%. This high percentage of accuracy is to be expected due to the huge imbalance between the number of valid and number of genuine transactions. Since the entire dataset consists of only two days' transaction records, its only a fraction of data that can be made available if this project were to be used on a commercial scale. Being based on machine learning algorithms, the program will only increase its efficiency over time as more data is put into it.

## REFERENCES

[1] "Credit Card Fraud Detection Based on Transaction Behaviour -by John Richard D. Kho, Larry A. Vea" published by Proc. of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017

[2] CLIFTON PHUA1, VINCENT LEE1, KATE SMITH1 & ROSS GAYLER2 " A Comprehensive Survey of Data Mining-based Fraud Detection Research" published by School of Business Systems, Faculty of Information Technology, Monash University, Wellington Road, Clayton, Victoria 3800, Australia

[3] "Survey Paper on Credit Card Fraud Detection by Suman" , Research Scholar, GJUS&T Hisar HCE, Sonepat published by International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 3, March 2014

[4] "Research on Credit Card Fraud Detection Model Based on Distance Sum – by Wen-Fang YU and Na Wang" published by 2009 International Joint Conference on Artificial Intelligence

[5] "Credit Card Fraud Detection through Parenclitic Network AnalysisBy Massimiliano Zanin, Miguel Romance, Regino Criado, and SantiagoMoral" published by Hindawi Complexity Volume 2018, Article ID 5764370, 9 pages

[6] "Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy" published by IEEE Transactions on Neural Networks and Learning Systems, VOL. 29, NO. 8, AUGUST 2018

[7] "Credit Card Fraud Detection-by Ishu Trivedi, Monika, Mrigya, Mridushi" published by International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 1, January 2016 [8] David J.Wetson,David J.Hand,M Adams,Whitrow and Piotr Jusczak "Plastic Card Fraud Detection using Peer Group Analysis" Springer, Issue 2008.

[8] Mubarek M and Adalı E., "Multilayer perceptron neural network technique for fraud detection," 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, 2017, pp. 383-387.

[9] Sahin Y and Duman E., "Detecting credit card fraud by ANN and logistic regression," 2011 International

Symposium on Innovations in Intelligent Systems and Applications, Istanbul, 2011, pp. 315-319.

[10] Saraswathi E., Kulkarni P., Khalil M. N and Chandra Nigam S., "Credit Card Fraud Prediction And Detection using Artificial Neural Network And Self-Organizing Maps," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 1124-1128.