International Journal of Marketing and Technology
Vol. 11 Issue 10, October 2021
ISSN: 2249-1058 Impact Factor: 6.559
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Evaluation of Hybridized Homogenous Supervised Learning Schemes in CreditCard Fraud Detection.

Ismaila W. Oladimeji[•]
Falohun Adeleye S[••]
Omotosho Oluyinka I.[•••]

Abstract

Fraud detection is a way of finding patterns in data that do not conform to expected behaviour. Fraud detection finds extensive use in a wide variety of applications such as fraud detection for credit cards, intrusion detection for cyber-security, military surveillance for enemy activities. As credit card being the primary method of payment in online transactions, credit card frauds have also been observed to surge as the number of online transactions have increased. The credit card industry has studied computing models for automated detection systems which have now been the subjects of academic research. This paper evaluates an ensemble homogenous supervised learning system (EHSLS) to detect fraud in credit cards online transactions. Random Forest is a type of supervised machine learning algorithm based on ensemble learning which is used for regression and classification tasks and well suited when the class distribution is unbalanced. The work flow of the proposed fraud detection system includes data preparation phase, implementation phase and evaluation phase. Cross validation technique was used for training and testing. The results showed the EHSLS produced 89.47%, 88.83%, 96.80%; CD-CPNN produced 93.80%, 91.70%, 95.13%; RBFN-PSO generated, 93.9%, 95.1% 91.7% while CPNN-GA gave 96.89%, 93.75%, 97.30% for recall , precision accuracy, respectively. However, the system developed produced the least f-measure value of 89.15%.

Keywords:

Fraud detection,
Credit cards, Cross validation, Random forest, F-measure

Author correspondence:

Ismaila W. Oladimeji,
Department of Computer Science,
LadokeAkintola University of Technology, Ogbomoso, Nigeria.

1. Introduction

Fraud detection refers to the problem of finding patterns in data that do not conform to expected behaviour. Many real-world applications such as intrusion or credit card fraud detection require an effective and efficient framework to identify deviated data instances. The techniques for finding these different deviations fall into three categories viz; (i) *Point anomalies* in which a point fraud is found by looking for specific individual samples in the data that are not similar to the rest of the data set. (ii)

Collective anomalies represent situations where the anomalous behaviour develops and extends over a number of data points that extend in time or space. and (iii) *Contextual Anomalies in which* ***a***nomalies are discovered when characteristics of the data are used to filter relevant data are termed "contextual" anomalies. The sources used to build context in the data are referred to as "contextual" attributes [18].

Fraud detection finds extensive use in a wide variety of applications such as fraud detection for credit cards, insurance, or health care, intrusion detection for cyber-security, fault detection in safety critical systems, and military surveillance for enemy activities [19]. The increasing popularity of e-commerce in daily lives has led to increase in credit card usages increased over the years. As credit card being the primary method of payment in online transactions, credit card frauds have also been observed to surge as the number of online transactions have increased [20]. Credit card fraud detection system is a computer program that attempts to perform fraud detection by identifying fraud or fraud transaction as quickly as possible once it has been perpetrated[21]. The credit card detection process is summarized in figure 1. The ultimate goal of such detection processes is to prevent the processing of all transactions that do not comply with the imposed regularities [22].

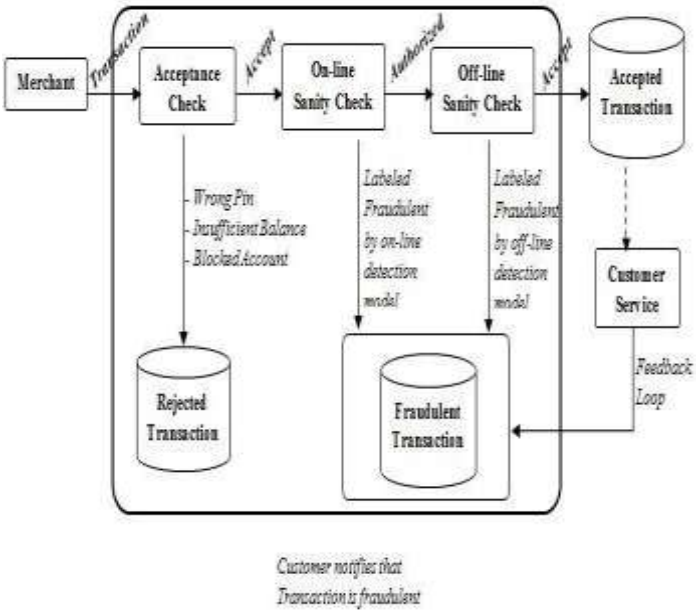

Figure 1: **Credit Card Detection Process [22]**

Some of the properties a fraud detection system should have in order to perform some good results [23] include but not limited to the following; the system should be able to handle skewed distributions, the ability to handle noise and overlapping of data, should be able to adapt themselves to new kinds of fraud, good metrics to evaluate the classified system, should take into account the cost of the fraudulent behavior detected and cost associated with stopping it.

For many years, the credit card industry has studied computing models (especially machine learning models) for automated detection systems which have now been the subjects of academic research [24]. In this work, the performance of an hybridized homogenous supervised learning scheme (Random forest) was evaluated in credit card fraud detection. Random forests are a scheme proposed by Leo Breiman in the 2000's for building a predictor ensemble with a set of decision trees that grow in randomly selected subspaces of data [10]. Random Forest could be a type of supervised machine learning algorithm based on ensemble learning. The Random Forest combines multiple algorithms of the same type i.e. multiple decision trees, leading to a forest of trees, thus the name "Random Forest". Random Forest is well suited when the class distribution is unbalanced. Random forest has the advantage over decision tree as it corrects the habit of over fitting to their training data sets. It has been found to provide a good estimate of generalization error and resistant to over fitting. The rest of the paper is organized as follow. Section 2 reviewed the related works on Random forest in fraud detection areas. Section 3 discussed briefly on random forest while section 4 gave a detailed steps of methodology employed. In section 5 is the discussion of results and finally section is conclusion.

2. RELATED WORKS

Over the years, the researchers have done a lot on finding the last solution to fraud problems in credit card industry especially in automated teller machines transactions. Their researches have based on

supervised and unsupervised learning, single and multiple algorithms, hybridized and ensemble schemes. Some of the works are highlighted as follow especially on random forest. The authors in [24]created a predictive model that capture the fraudulent transactions with high accuracy using Isolation forest and Local Outlier factor for detecting outliers that explicitly identifies anomalies and Extreme Gradient Boosting, an ensemble approach for constructing and evaluating the predictive model. A comparative study was done with existing models Logistic Regression, SVM, Random Forest with Extreme Gradient Boosting algorithm. The proposed model showed better performance and secured high accuracy of 0.98. In 2019, [3] proposed a system where Random Forest Algorithm was used for classification and regression. In credit card fraud detection, credit card data sets were collected for trained data sets and user credit card queries are collected for testing data sets. After classification process, Random Forest Algorithm was used for analysing data sets and current data sets. Finally, the optimization was done and the accuracy obtained by Random Forest was 99.9%

In 2015, the authors in [8] proposed a novel technique for generating Random Forest by calculating a weighted group of predictive probabilities and then taking random samples of many trees from earlier distributions. This technique uses power likelihood instead of likelihood, which decides space spanned by the combination of trees. It is called safe – Bayesian because even though the underlying probabilistic model is wrong, it gives good predictive performance. They have proved using nine different datasets that the proposed technique of Safe – Bayesian Random Forest gives better results than classification algorithms like K nearest Neighbors. The researchers in [9] proposed two approaches based on Random Forest to achieve improved generalization in the analysis of hyper spectral data, when the volume of training data is small.

The new classifier is suggested with Bagging of training samples and Adaptive random subspace feature selection within binary hierarchical classifier. This caused the number of features selected at each node to be dependent on quantity of relevant training data. The results showed that RF-BHC proved to be superior than RF-CART. [5] used standard scalar model to identify whether a new transaction is fraudulent or not. The trained standard scalar model with high probability considered an incoming transaction to be fraudulent and not acceptable. Thus, the Random forest built multiple decision trees and integrate them together to get stable prediction and accuracy of about 98.6%.

In 2018, the authors [6] discussed a Big data analytical framework to process large volume of data and implemented various machine learning algorithms (decision tree and logistic regression) for fraud detection and observed their performance on benchmark dataset to detect frauds on real time basis there by giving low risk and high customer satisfaction. Also in 2015, the researchers in [1] introduced Random Forest for financial fraud technique detection and detailed features selection, variables' importance measurement, partial correlation analysis and Multidimensional analysis. The results show that a combination of eight variables has the highest accuracy. Moreover, four statistic methodologies were applied including Random Forest which has the highest accuracy. **I**n the work of [2], they used machine learning algorithms to detect credit card fraud. Standard models and then employed hybrid methods which use AdaBoost and majority voting methods. To evaluate the system a publicly available credit card data set with noise added. The results indicated that the majority voting method achieves good accuracy rates in detecting fraud cases in credit cards.

The authors in [16] experimented the performance of optimization of hybridized counter propagation neural network (CPNN) with genetic algorithm (GA) to detect anomaly in credit cards online transactions. CPNN-GA anomaly detection system gave 97.3%, 0.0%, 3.7% and 2.0% for prediction accuracy, false acceptance rate, false rejection rate and equal error rate respectively. The false alarm rate for the GA, CPNN and CPNN-GA are 0.83%, 1.35% and 0.61% respectively. [17] employed Communal detection (CD) and Counter Propagation Neural Network (CPNN) for fraud detection in identity theft in credit cards online transactions. The selected simulated applicants attributes were worked on by CD to produce whitelists and blacklists. The result- cum-applicants data were preprocessed and passed to the CPNN to performed classifications.

The results showed that CD-CPNN system produced average false positive rate, average false alarm rate, average detection rate and average prediction accuracy of 26.3, 2.7, 93.6, and 92.5% respectively. Ismaila and Ismaila (2019c) investigated the extent to which Radial Basis Function (RBFN)-cum-Particle Swarm Optimization (PSO) was used to detect frauds in credit card online transactions. The simulated dataset used contained legal transactions sparsely intertwined with malicious types. The results showed that RBFN-PSO generated 95.1%, 23.0%, 91.7% and 93.9% for accuracy, false alarm rate, precision and recall respectively. However, ensemble homogenous algorithm like Random forest has not been compared with hybridized heterogeneous based fraud detection systems for performance.

3. RANDOM FOREST

To derive any Random forest, Decision Tree is the rudimentary. Random Forest is basically composed of simple tree predictors. Random forest is best suited for large datasets and at the same time the learning algorithm produces accurate results and handles missing data and exhibits good performance results.. [10]. Random Forest is well suited when the class distribution is unbalanced [11]. Single-Tree model such as decision tree are very sensitive to specific training data and more prone to over fitting of data [12]. By using ensemble methods somehow over fitting problem can be reduced by combining a group of decisions [13] thereby improving the accuracy of results.

The accuracy of random forest depends upon accuracy of each individual tree as well as correlation between the groups of trees. The better accuracy of each individual tree collectively will give the best performance results for the ensemble tree. The variation and their randomness of a tree will usually come by selecting different subsets of attributes during the construction of decision tree. The training set for each and every individual decision tree is a group of randomly chosen training data. At every internal node of the tree, it again randomly selects some subset of attributes and then computes the center.

The LeftCenter and RightCenter are denoted as Class 0 and Class 1. The kth element of a center is calculated (Abeel and Saeys, 2009) with the below formulas

$$LeftCenter[k^{th}\ \text{element}] = \frac{1}{n}\left(x_{ik} I\left(y = 0\right)\right) \quad \text{.....(1)}$$

$$RightCenter[k^{th}\ \text{element}] = \frac{1}{n}\left(x_{ik} I\left(y = 1\right)\right) \quad \text{.....(2)}$$

In the present node, each of the element in the train data set is classified by calculating Manhattan Distance between the element and the center of the node, it is calculated as:

$$\text{distance(center, element)} = \sum_{i \in sub} |\,(center[i] - i^{th}\ element) \quad \text{.....(3)}$$

Sub is the sub-set of attributes which are randomly selected from the given set of attributes (X).

4. METHODOLOGY

The work flow of the proposed ensemble homogenous supervised scheme based fraud detection system include data preparation phase, implementation phase and evaluation phase. The workflow is shown in figure 2.

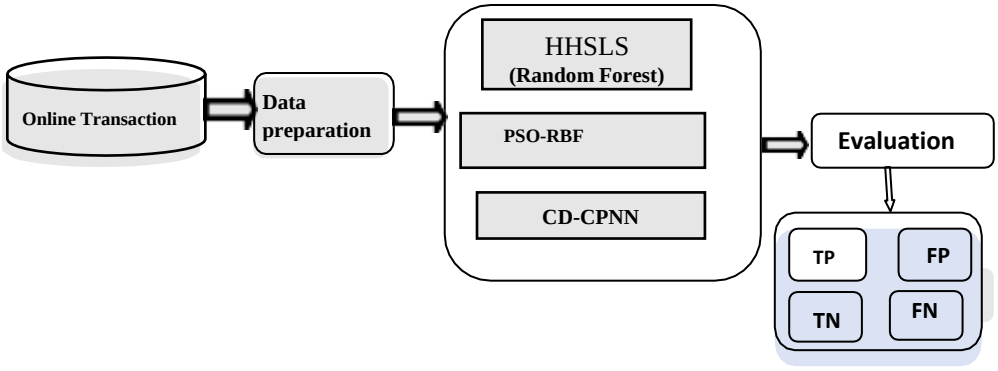

Figure 2: Work flow of the proposed EHSLS system

Data preparation

A simulator was used to generate a mix of genuine and fraudulent (sparsely inter-twined) transactions. This stage involves preparing the generated data for training. It is data mapping phase which has to do with matching the parameters of Random forest to selected variables of generated cardholder's

transaction data. Also, Several attributes that are ordered categorical have been coded as integer, for instance the predicted response class label y, was dichotomously defined as follows:

$$y = W(x) = \begin{cases} 0, & if\ a\ transaction\ is\ non-fraud \\ 1, & if\ s\ transaction\ is\ fraud \end{cases} \quad \ldots(4)$$

Implementation Phase

This phase encompassed the training phase and prediction phase. The prediction phase employed the parameters settings from trained Random forest model to identify the new transactions as fraudulent or not. The pseudocode of the Random Forest is shown in the figure 3. This random forest technique was adopted to learn responses like "fraud" and "not fraud". When the data sample will be fraudulent, expected response of the Random forest was 1, and produced 0 otherwise

.

Step 1: Let D = {($x1$, $y1$), . . . , (xN, yN)} denote the training data,with xi =(xi,1, . . . , xi,p)T . For j = 1 to J:

Step 2. Take a bootstrap sample Dj of size N from D.

Step 3: Using the bootstrap sample Dj as the training data, fit a treeusing binary recursive partitioning:

(a) Start with all observations in a single node.
(b) Repeat the following steps recursively for each un-splitnode until the stopping criterion is met:
(i). Select m predictors at random from the p availablepredictors.
(ii). Find the best binary split among all binary splits onthe m predictors from step i.
iii. Split the node into two descendant nodes using thesplit from step ii.

Step 4: To make a prediction at a new point x,f (x) = 1$J\Sigma^{J}_{j=1}h_j(x)$ for regression
f (x) = argmax $y\Sigma^{J}_{j=1}I(h_j(x) = y)$ for classification where $\hat{h}\,j(x)$ is the prediction of the response variableat x using the jth tree.

Figure 3: Pseudocode of the Random Forest

Evaluation Metrics

In this work, confusion matrix parameters were employed for evaluating the results of the developed system, the metrics used are viz; accuracy, precision, recall and f-measure.

$$\text{Recall} = \frac{TP}{TP+FN}\ x\ 100 \quad \ldots\ldots\ldots(4)$$

$$Precision = \frac{FP}{TP+FP}\ x\ 100 \ldots\ldots(5)$$

$$Accuracy = \frac{TP+TN}{TP+FN+TN+FP}\ x\ 100 \quad \ldots\ldots(6)$$

$$\text{F-Measure} = \frac{2 \text{ x (Precision x Recall)}}{\text{(Precision + Recall)}} \quad \ldots(7)$$

Where FP-false positive TP-true positive
FN-false negative TN-true negative

5. RESULTS AND DISCUSSIONS

The interface of the developed fraud system was done using Matrix Laboratory (MATLAB). as shown in figure 3, 4 and 5. Figure 3 shows the interface for card transactions, figure 4 depicts the training of data by the system, while figure 5 shows the prediction scene of the system. A simulator was used to generate a total of 1,300 data

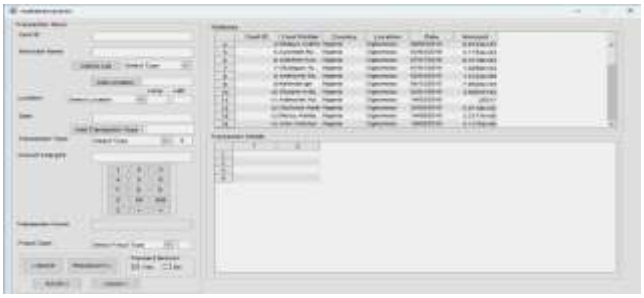

Figure 3. Graphical Interface for Card transactions

intertwined of genuine and fraudulent transactions. Cross validation method was used to carry out training and testing stages. The experiments were performed using fifty percent training and testing

Figure 4: Matlab GUI for developed system

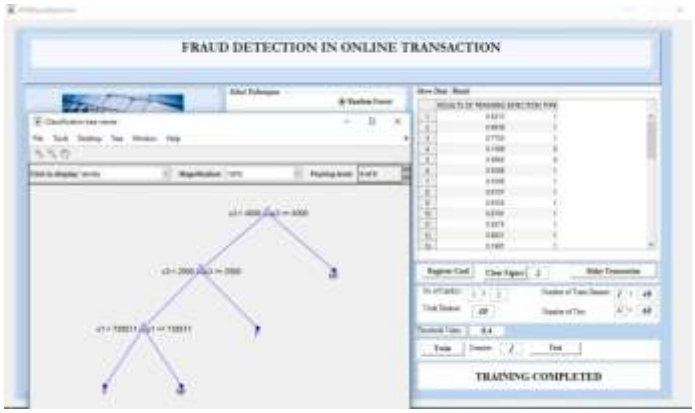

Figure 5.Matlab GUI Prediction stage

The results of the developed fraud detection system were shown in table 1 with comparison with heterogeneous hybridized schemes. The system developed produced 89.47%, 88.83%, 96.80%; CD-CPNN produced 93.80%, 91.70%, 95.13%; RBFN-PSO generated, 93.90%, 95.10% 91.70% while CPNN-GA gave 96.89%, 93.75%, 97.30% for recall , precision accuracy, respectively. However, the system developed produced the least f- measure value of 89.15%.

Table 1: The results of System developed in Comparison with others

	System developed	CD-CPNN	PSO_RBN	CPNN-GA
Recall	89.47	93.80	93.90	96.89
Precision	88.83	91.73	91.70	93.75
Accuracy	96.80	95.13	95.10	97.30
F-measure	89.15	92.75	92.78	95.29

6. CONCLUSION

The development of fraud detection system using Random Forest to classify transactions types (fraudulent and non-fraudulent) was done. Random Forest is basically composed of simple tree predictors. Random forest is best suited for large datasets and at the same time the learning algorithm produces accurate results and handles missing data and exhibits good performance results. Random forest algorithm or classifier can be used for Classification as well as Regression tasks. It efficiently handles missing data and won't over fit the model if it there are more trees. A dataset comprising three thousand transactions (genuine and fraudulent) were simulated. The dataset were mapped with the parameters of Random forest for processing. The evaluation results showed that the RF based system produced 89.47%, 88.83%, 96.80%, 89.15% for recall , precision accuracy, f-measure value respectively. However, ensemble homogenous algorithm like Random forest being compared with hybridized heterogeneous based fraud detection systems showed that it did not perform better. However, the developed system can be improved by using features extraction scheme to remove redundant features.

REFERENCE

[1] Chengwei Liu, Yixiang Chan, Syed Hasnain Alam Kazmi1 & Hao Fu (2015).Financial Fraud Detection Model: Based on Random Forest, of Economics and Finance; Vol. 7, No. 7; 2015 ISSN1916-971X E-ISSN 1916-9728

[2] Siva Prakash S., Ahubhakumar, Appash S., Cibiragul J. (2019). Credit Card Fraud Detection using Adaboost and Majority Voting, International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, RTICCT - 2019 Conference Proceedings, Vol. 7, Issue 01, pp 1-4.

{3] Monika S., Venkataramanamma K., Pritto Paul P., Usha M. (2019). Credit Card Fraud Detection using Random Forest Algorithm, International Journal of Research in Engineering, Science and Management Volume-2, Issue-3, March-2019, pp 131-133.

[4] Chandra Sekhar Kolli, T.Uma Devi (2019). Isolation Forest and Xg Boosting For Classifying Credit Card Fraudulent Transactions, International Journal of Innovative Technology and ExploringEngineering (IJITEE) ISSN: 2278-3075, volume-8 Issue-8 June, 2019.

[5] Niveditha G., Abarna K., Akshaya G. V. (2019) Credit Card Fraud Detection Using Random Forest Algorithm, International Journal of Scientific Research in Computer Science, Engineering and Information Technology © 2019 IJSRCSEIT, vol. 5, Issue 2.

[6] Patil S., Nemade V., Soni P. (2018). Predictive Modelling For Credit Card Fraud Detection Using Data Analytics, International Conference on Computational Intelligence and Data Science (ICCIDS 2018), Procedia Computer Science 132 (2018) 385–395.

[7] Rashmi H Roplekar, Buradkar N. V. (2017). Survey of Random Forest Based Network Anomaly Detection Systems, International Journal of Advanced Research in Computer and Communication Engineering, vol. 6, Issue 12.

[8] Quadrianto, N. and Ghahramani, Z. (2015). A Very Simple Safe-Bayesian Random Forest. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6), pp.1297-1303.

[9] Ham, J., Yangchi Chen, Crawford, M. and Ghosh, J. (2005). Investigation of the random forest framework for classification of hyperspectral data. *IEEE Transactions on Geoscience and RemoteSensing*, 43(3), pp.492-501.

[10] Breiman L, Random forests. Machine Learning, 5-32, and 2001.

[11] Bhattacharyya S, Jha S, Tharakunnel K, and Westland J.C,(2011). "Data mining for credit cardfraud ,a comparative Study", Decision Support Systems 50, 602-13.

[13] Khoshgoftaar T. M, Golawala M and Van Hulse J, "An Empirical Study of Learning from Imbalanced Data Using Random Forest", in 19th IEEE International Conference on Tools with AI PP310–317, 2007.

[14] Dietterich T.G, "Ensemble methods in machine learning. 1-15, 2000.

[15] Abeel T, and Saeys Y, "A Machine Learning Library, Journal of Machine Learning Research", 931-934, 2009.

[16] Ismaila W. Oladimeji, Ayodele A. Lawrence, Omidiora E. Olusayo, Falohun A. Samuel. (2019). Soft Computing: Optimization of Hybridized Neural Network Variant with Genetic Algorithm for Anomaly Detection, International Journal Of Innovative Research & Development (IJIRD), Vol 8 Issue 9, pp. 250- 257.

[17] Ismaila W. Oladimeji, Ismaila Folasade. M., Falohun Adeleye S. (2020). Investigation of Two-Step Approach to Identity Theft Detection in Electronic Payments**,** International Journal of Engineering & Scientific Research. (manuscript accepted).

[18] Aleskerov, E., Freisleben, B., and Rao, B. (1997). Cardwatch: a nueral network based database minining system for Credit card fraud detection. In proceedings of the IEEE Conference on Computational Intelligence for Financial Engineering, 220-226.

[19] Chandola V., Eilertson, E., Ertoz, L., Simon, G., and Kumar, V. (2006). Data Mining for Cyber Security. In Data Warehousing and data Mining Techniques for Computer Security, A. Singhal, ed. Springer.

[20] Minyong L., Seunghee H., and Qiyi Jiang (2013). E-commerce transaction anormaly classification.

[21] Falaki S. O., Ismaila W. O., Ayeni J. O., Alese B. K., Adewale O. S., Aderounmu G. A (2012). Effect of Hybrid Hidden Markov Models in Fraud Detection System. International Journal of Scientific Innovation And Sustainable Development, Ghana.

Eliassi-Rad

[22] Vlasselaer V.C., Bravo C., Caelen O., Eliassi-Rad, Akoglu L., Snoeck M., and Baesens B. (2015). APATE: A novel Approach for automated credit card transaction fraud detection using network-based extensions. Decision Support Systems.

[23] Nimisha Philip, Sherly K.K (2012). Credit Card Fraud Detection Based On Behavior Mining, TIST.Int.J.Sci.Tech.Res.,Vol.1 (2012), 7-12.

[24] Chan P. K., Fan W. Prodromidis A. L. and Stolfo S. J. (1999). Distributed Data Mining in credit card fraud detection, *IEEE Intelligent Systems*, pp 67-74.