International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 9 ǁ September 2022 ǁ PP. 97-104

A Novel Credit Card Fraud DetectionBased On Ensemble Learning Algorithms

Dr. S.Naveen Kumar[1], K. Jaswanth[2]

[1] *Associate Professor, Dept of CSE, Audisankara College of Engineering andTechnology (AUTONOMOUS), Gudur, AP, India.*

[2,] *PG Scholar, Dept of MCA, Audisankara College of Engineering andTechnology(AUTONOMOUS), Gudur, AP, India.*

ABSTRACT

As technology developed, new business-making mechanisms emerged in the financial sector. One of them is the credit card system. But due to several flaws in this method, numerous issues are raised in this system by credit card frauds. The industry as well as customers who use credit cards is suffering greatly as a result. Lessons on investigating actual credit card figures in relation to privacy concerns are lacking. In the publication, an effort has been made to uncover credit card fraud using algorithms that used machine learning approaches. . In this regard, two algorithms are used via Fraud Detection in credit card using Decision Tree and Fraud Detection using Random Forest. The efficiency of the model can be decided by using some public data as sample. Then, an actual world credit card facts group from a financial institution is examined. Additionally, additional noise is added to the data samples in order to auxiliary assess the systems' durability. The first approach in the study is significant since it builds a tree against the user's behaviors, and by utilizing this tree, frauds will be detected. In the second way, a user activity-based forest will be built, and it will be attempted to identify the suspect using this forest. The findings of the analysis unequivocally demonstrate that the common elective method detects credit card fraud situations with respectable degrees of precision.

Keywords: Gaussian Mixture, Bayesian Network, Clustering, DAG, Credit Card,Tree, Forest, Scam.

Date of Submission: 25-08-2022 Date of acceptance: 09-09-2022

I. INTRODUCTION

When a firm takes precautions against whipped cash, goods, or amenities obtained through an unlawful credit card operation, credit card scam detection is taking place. Customers and third parties can both be the victims of credit card fraud. There are many techniques developed to prevent such frauds. If such frauds occur, then methods for locating the improper transactions are also devised. Many new and original algorithms have been put out to protect digital data transfers against unwanted access. However, there are certain negative aspects in one way or another. This essay discusses techniques for identifying credit card fraud.

For the purpose of detecting fraud, a variety of supervised and semi-supervised machine learning techniques are used. However, our goal is to address three key issues with the card fraud dataset, namely, the strong class imbalance, the inclusion of labeled and unlabelled samples, and the need to process a large volume of transactions**.**

Different To identify fraudulent transactions in real-time datasets, supervised machine learning algorithms such as Decision Trees, Naive Bayer’s Classification, Least Squares Regression, Logistic Regression, and random forest algorithm are utilized. To train the behavioral characteristics of typical and aberrant transactions, two random forest techniques are utilized. They are CART-based and Random-tree-based random forests. Despite the fact that random forest produces decent results on despite the tiny data set, there are still some issues when the data is unbalanced. The upcoming work will concentrate on resolving the aforementioned issue. The random forest algorithm itself needs to be improved.

Research is being done on investigating meta-classifiers and meta-learning methodologies in managing highly skewed credit card fraud data in order to study the performance of Logistic Regression, K-Nearest Neighbor, and Nave Bayes. Using supervised learning techniques to identify fraud instances may not always be successful. a deep auto-encoder and restricted Boltzmann machine (RBM) model that may create typical transactions to identify abnormalities in typically occurring patterns. Additionally, a hybrid technique that combines the Adaboost and Majority Voting procedures has been devised.

II. LITERATURE SURVEY

Multiple Supervised and Semi-Supervised machine learning techniques are used for fraud detection [8], but we aim is to overcome three main challenges with card frauds related dataset i.e., strong class imbalance, the inclusion of labelled and unlabelled samples, and to increase the ability to process a large number of transactions.

Different Supervised machine learning algorithms [3] like Decision Trees, Naive Bayes Classification, Least Squares Regression, Logistic Regression and SVM are used to detect fraudulent transactions in real-time datasets. Two methods under random forests [6] are used to train the behavioral features of normal and abnormal transactions. They are Random-tree-based random forest and CART-based. Even though random forest obtains good results on small set data, there are still some problems in case of imbalanced data. The future work will focus on solving the above-mentioned problem. The algorithm of the random forest itself should be improved.

Performance of Logistic Regression, K-Nearest Neighbor, and Naïve Bays are analyzed on highly skewed credit card fraud data where Research is carried out on examining meta-classifiers and meta-learning approaches in handling highly imbalanced credit card fraud data.

Through supervised learning methods can be used there may fail at certain cases of detecting the fraud cases. A model of deep Auto-encoder and restricted Boltzmann machine (RBM) [2] that can construct normal transactions to find anomalies from normal patterns. Not only that a hybrid method is developed with a combination of Adaboostand Majority Voting methods [4].

III. PROPOSED ANALYSIS

When contrasted to the customer's prior purchases, card transactions are always foreign. This When they are known as idea drift difficulties, unfamiliarity is a particularly challenging problem in the actual world. It is possible to think off concept drift as a variable that evolves over time and in unexpected ways. These factors significantly unbalance the data. Our research's primary goal is to find a solution to the Concept Drift issue for real-world application.

Attribute name	Description
Transaction id	Identification number of transaction
Cardholder id	Unique identification number given to the cardholder
Amount	Amount transferred or credit in a particular transaction by the customer
Time	Details like time and date, to identify when the transaction was made
Label	To specify whether the transaction is genuine or fraudulent

Table 1: Raw features of credit card transactions

Table 1, [1] shows basic features that are captured when any transaction is made.

Dataset Description

The dataset [11] contains transactions made by a cardholder in duration in 2 days i.e., two days in the month of September 2013. Where there are total 1,00,006 transactions among which there are 492 i.e., 0.172% transactions are fraudulent transactions. This dataset is highly unbalanced. Since providing transaction details of a customer is considered to issue related to confidentiality, therefore most of the features in the dataset are transformed using principal component analysis (PCA). V1, V2, V3,..., V28 are PCA applied features and rest i.e., ‘time’, ‘amount’ and ‘class’ are non-PCA applied features, as shown in table 2

S.no	**Feature**	**Description**
1.	Time	Time in seconds to specify the elapses between the current transaction and first transaction.
2.	Amount	Transaction amount
3.	Class	0 – not fraud 1 – fraud

Table 2: Attributes of European dataset

According to this matrix, the attribute class is unrelated to the transaction's value and timing. Even from the matrix, it is evident that the qualities used in PCA determine the transaction's class.

Algorithm: Algorithm to derive aggregated transaction details and to extract card holder features using sliding window technique

```
l: length of TGenuine= [];Fraud= [];
For i in range 0 to l-w+1:
T: [];
/* sliding window features*/For j in range i+w-1:
/*Add the transaction to window */T=T+tj id;
End
/* features extraction related to amount */ai1=MAX_AMT(Ti); ai2=MIN_AMT(Ti); ai3=AVG_AMT(Ti);
ai4=AMT(Ti);
For j in range i+w-1:

End

/* Time elapse */
xi= Time(tj)-Time(tj-1)

Xi= (ai1, ai2,ai3,ai4,ai5,);Y= LABEL(Ti);
/* classifying a transaction into fraud or not */if Yi=0 then

Else

End

Genuine =Genuine U Xi;

Fraud =Fraud U Xi;
```

The old transactions are eliminated when a new one is fed to the window, and step 2 is carriedout for each set of transactions. (The Sliding-Window based approach of aggregation algorithm is referred from. After pre-processing, we use the cardholders' behavior patterns in each group totrain several classifiers and extract fraud characteristics. Even when we apply classifiers to the dataset, they do not perform well because of the imbalance (shown in fig. 4) in the dataset.

Decision Tree based Credit Card Fraud Detection Algorithm using Machine Learning.

When it is necessary to aggregate the unusual events in a business from a recognised client, the algorithm is applied. It is one of the statistical predictive modelling strategies. This algorithm's enforced evaluation of all potential outcomes of decisions, tracking of all possible routes to a conclusion, and thorough study of the results are some of its key advantages.

Use Case: A situation wherever a customer makes businesses is considered. The decision tree is constructed to forecast the possibility of scam centered on the business made as shownin figure-1.

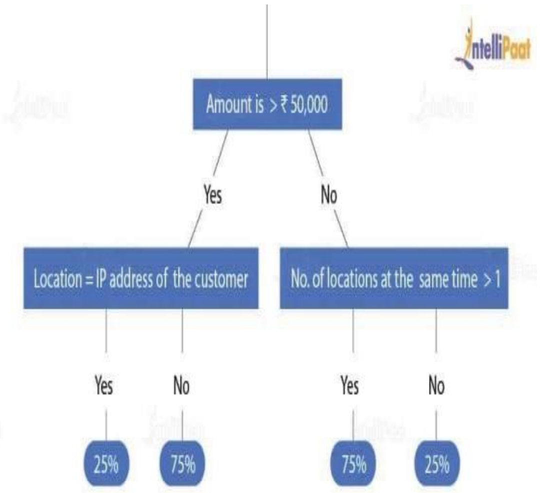

Amount is > ₹ 50,000
Yes
No
Location = IP address of the customer
No. of locations at the same time > 1
Yes
No
Yes
No
25%
75%
75%
25%

Fig. 1: User Transactions Tree

Random Forest based Credit Card Fraud Detection Algorithm using MachineLearning.

This technique, which employs a mixture of decision trees to provide a better result, is an upgraded version of the decision tree algorithm. Any single decision tree design for any circumstance will operate on any piece of data and on any decision tree. Every tree has the potential for both legitimate and fraudulent enterprise.

Random decision forests and Random forests are group learning techniques for categorization, prediction, and other tasks. They work by constructing a massive volume of decision trees during training time and producing the class that is the mean prediction (regression) or mode of the modules (categorization) of the individual trees. Random decision forests exact for decision treesnature of over fitting to their exercise set.

Use Case: Consider a scenario where a transaction is made. Now, an illustration is made on the way the random forest in Machine Learning is used in scam finding algorithms is as shown in figure-2.

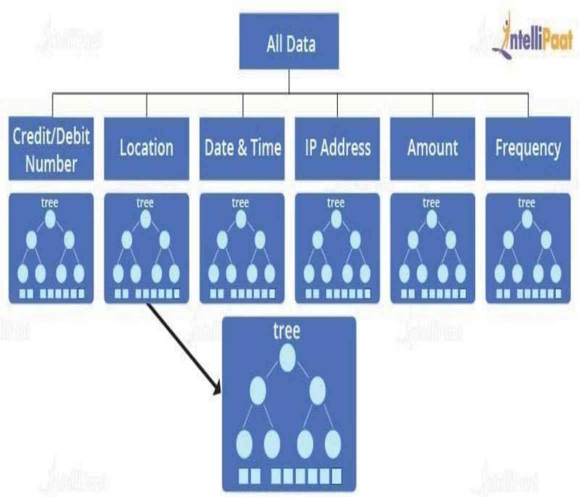


Fig. 2: User Transaction Forest

An approach used in ensembles is random forest. As a result of its simplicity and diversity, it is one of the most used algorithms. Many decision trees are used in this model. Each of these decision trees separates a class of predictions, and the group receiving the most votes becomes the final output prediction of our model, as illustrated in fig. 4. When developing trees in a random forest, it seeks for the best characteristics from a random subset of traits to split the nodes, rather than the most crucial features. This results in a great deal of variability, which will improve our model. The models provide ensemble forecasts that are more accurate than any of the individual predictions since there is very little connection between the many models that are generated. This is due to the possibility of incorrect trees.

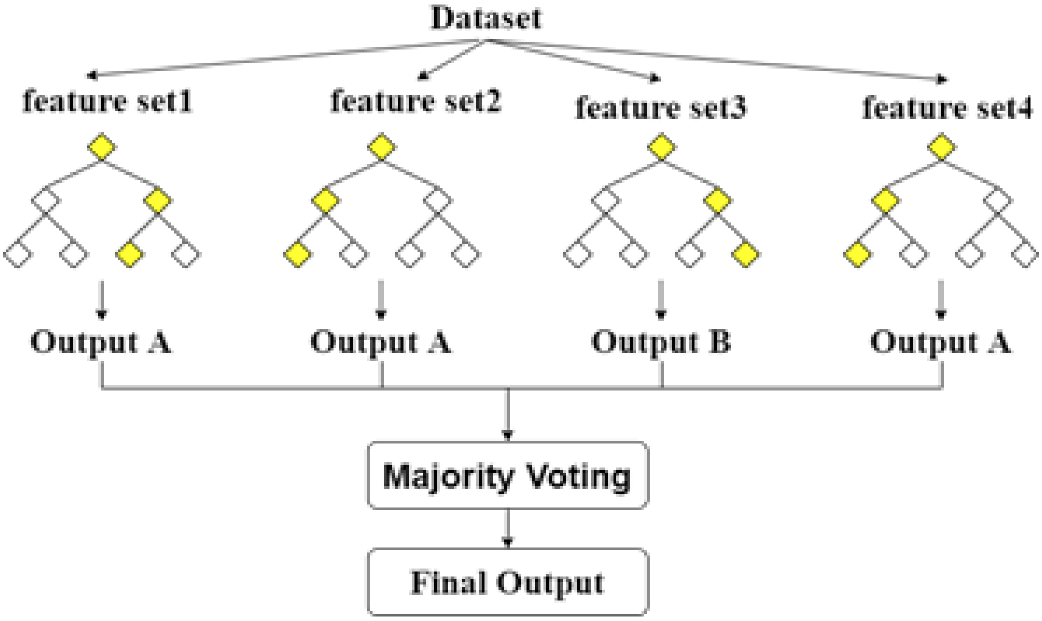


Fig. 3: Shows the correlation matrix of the dataset.

Algorithm steps are

In this algorithm first, random samples are selected from the given dataset.The algorithm creates decision trees for each of the samples selected.

Algorithm gets the predicted outputs to form each ofthe decision trees.

Voting is then performed for all the predicted outputs which were obtained from the decision trees. Finally, the result which gets the highest number of votes will be declared as the final predicted result.

Proposed Methodology

This section outlines the steps involved in holding a credit card hostage. Credit card companies use a variety of effective techniques to identify and stop frauds, including arrangement orientation, device learning, neural networks, artificial intelligence, and fuzzy logic. In recent years, credit card theft has grown increasingly prevalent. . In Current day, the fraud is one of the key causes of excessive business losses, not only for merchants, distinct clients are also affected. So there are some methods to detect such kind of frauds. Initially, clustering model was adopted to categorize the authorized and deceitful operation by means of data cauterization of areas of factor value. Furthermore, Gaussian mixture model is used to model the possibility thickness of credit card operator's past performance such that the chance of present actions can be intended to perceive any irregularities from the historical behavior. . Finally, the measurements of a particular user and the indicators of various fraud scenarios are defined using Bayesian networks. Figure 3 below displays an illustration of thesuggested model.

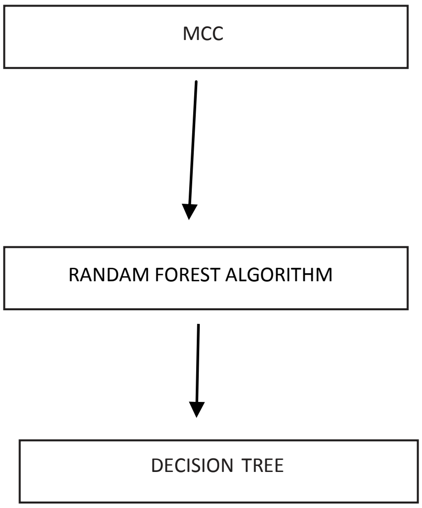


Fig. 4: Proposed Methodology

In our proposed system we use the following formulae to evaluate, accuracy and precision are never good parameters for evaluating a model. But accuracy and precision are always considered as the base parameter to evaluate any model.

The Matthews Correlation Coefficient (MCC) is a machine learning measure which is used to check the balance of the binary (two-class) classifiers. It takes into account all the true and false values that is why it is generallyregarded as a balanced measure which can be used even if there are different classes.

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

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{MCC} = \frac{TP*TN-FP*FN}{\sqrt{(TP+FN)(TP+FN)(TN+FN)(TN+FN)}}$$

TP –True Positive TN- True NegativeFP- False Positive FN- False Negative

IV. EXPERIMENTAL RESULTS AND ANALYSIS

On the original dataset as well as the SMOTE dataset, we have tested a few models. The data are summarized, and the accuracy, precision, and MCC all display significant variations. Even better for decision tree class datasets, we employed the one-class random forest algorithm. Our dataset contains two classes, thus we can also utilize the one-class random forest algorithm.

Methods	Accuracy	precision	Random forest
Local outlier forest	0.8990	0.0038	0.0172
Isolation forest	0.9011	0.0147	0.1047
Support vector machine	0.9987	0.7681	0.5257
Logistic regression	0.9990	0.875	0.6766
Decision tree	0.9994	0.8854	0.8356
Random forest	0.9994	0.9310	0.8268

Table 3: Accuracy, Precision and MCC values before applying SMOTE

Table 3, shows the results on the dataset before applying SMOTE and fig 5, shows the sameresults graphically.

One-Class Random Forest

Accuracy: 0.7009 Precision: 0.7015

Methods	Accuracy	Precision	MCC
Local outlier factor	0.4582	0.2941	0.1376
Isolation forest	0.5883	0.9447	0.2961
Logistic regression	0.9718	0.9831	0.9438
Decision tree	0.9708	0.9814	0.9420
Random forest	0.9998	0.9996	0.9996

Table 4: Accuracy, Precision and MCC values after applying SMOTE

Table 4, shows the results on the dataset after applying SMOTE and fig 6, shows the sameresults graphically.
By using our proposed system we will get below outputs, In the below figure, we are calculatethe accuracy levels in the form of pie charts.

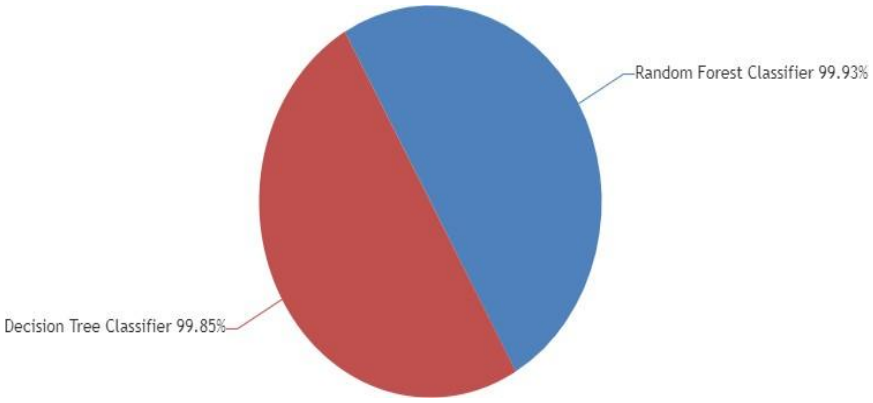


Fig4: accuracy results in the form of pie chart

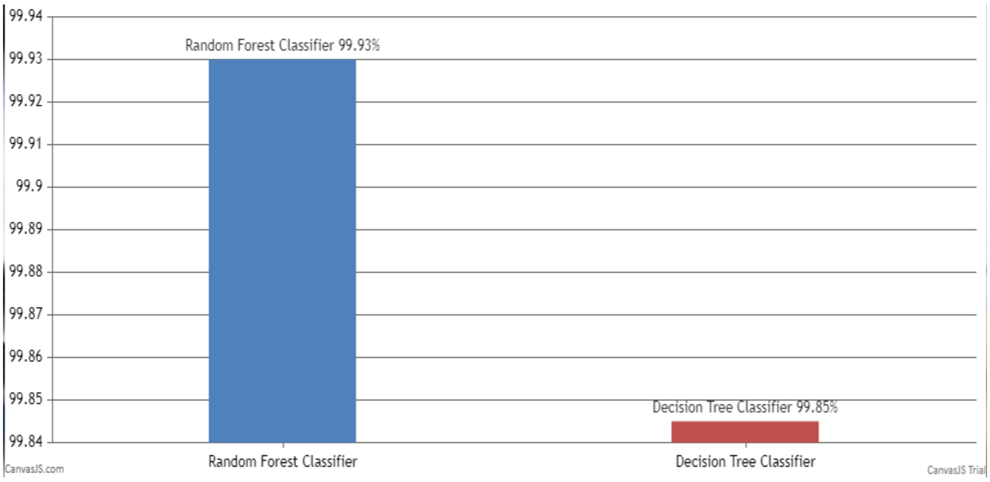
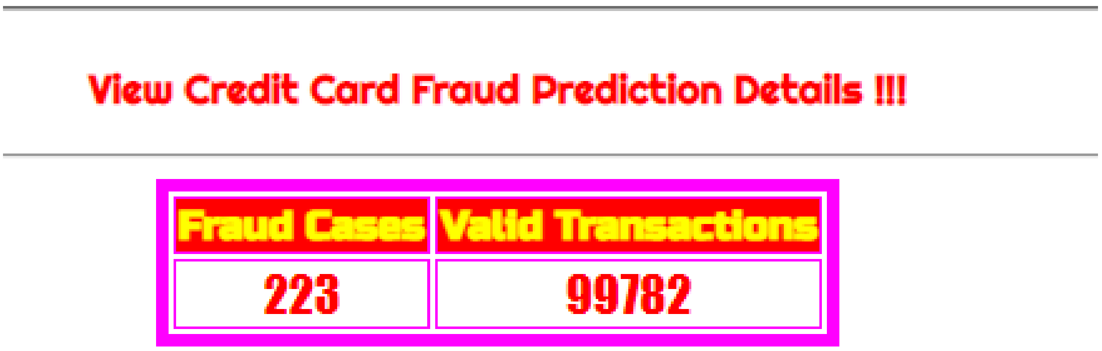


Fig5: Accuracy values in bar charts

In the below figure , Here we are able to see Accuracy values in bar charts
In the below figure, display the output based on the given data set and it displays the credit cardfraud cases and valid transactions



View Credit Card Fraud Prediction Details !!!

Fraud Cases	Valid Transactions
223	99782

Fig. 6: Detected Fraud Data values

V. CONCLUSION

In this study, we created a unique fraud detection technique that groups clients according to their transactions. and analyze behavior to create a profile for each cardholder. Following the application of various classifiers to three distinct groups, rating scores are produced for each type of classifier. The system adapts as a result of these dynamic changes in the parameters. Prompt response to new cardholder's transactional behaviors. A feedback system is then used to address the issue of notion drift. We The Matthews Correlation Coefficient was shown to be the superior metric for handling imbalance datasets. It wasn't only MCC. solution. We attempted to balance the dataset by using SMOTE and discovered that the classifiers were performing better than before. The use of one-class classifiers, such as one-class SVM, is an alternative method for addressing imbalance datasets. Finally, we found that the algorithms that produced the best outcomes were random forest, decision tree, and logistic regression.

REFERENCES

[1]. Adewumi and A. A. Akinyelu, "A survey of machine learning and nature-inspired based credit card fraud detection techniques," International Journal of System Assurance Engineering and Management, vol. 8, pp. 937– 953, 2017J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[2]. A. Srivastava, A. Kundu, S. Sural, A. Majumdar, "Credit card fraud detection using hidden Markov model," IEEE Transactions on Dependable and Secure Computing, vol. 5, no. 1, pp. 37–48, 2008K. Elissa, "Title of paper if known," unpublished.

[3]. Bansal, J. C., Singh, P. K., Saraswat, M., Verma, A., Jadon, S. S., and Abraham, A. (2011). Inertia weight strategies in particle swarmoptimization. In Nature and Biologically Inspired Computing (NaBIC), (Salamanca, Spain, October 19 - 21, 2011).IEEENaBIC'11,633--640.

[4]. Bello - Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. Information Fusion. 28 (Mar.2016), 45--59

[5]. Bharill, N., Tiwari, A., and Malviya, A. (2016). Fuzzy Based Clustering Algorithms to Handle Big Data with Implementation on Apache Spark.In Proceedings of the IEEE 2nd International Conference on Big Data Computing Service and Applications, (Oxford, UK, March 29-April 01, 2016). IEEE BigDataService '16, 95--104.

[6]. Y. Sahin, S. Bulkan, and E. Duman, "A cost -sensitive decision tree approach for fraud detection," Expert Systems with Applications, vol. 40, no. 15, pp. 5916–5923, 2013.

[7]. TheNilsonReport(October2016)[Online].Available:https://www.nilsonreport.com/upload/content_promo/The_ Nilson _Report_10-17-2016.pdf

[8]. J. T. Quah, and M. Sriganesh, “Real-time credit card fraud detection using computational intelligence,” Expert Systems with Applications, vol. 35, no. 4, pp. 1721–1732, 2008.

[9]. S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C., “Data mining for credit card fraud: A comparative study,” Decision Support Systems, vol. 50, no. 3, pp. 602–613, 2011.

[10]. S. Panigrahi, A. Kund u, S. Sural, and A. K Majumbar, “Use of Dempster-Shafer theory and Bayesian inferencing for fraud detection in communication networks”, Lecture Notes in Computer Science, Spring Berlin/ Heidelberg, Vol. 4586, , 2007, p.446-460