ANALYTICAL MEASURES FOR DETECTING FRAUD USING CLASSIFICATION ALGORITHMS

D .Vimal Kumar[1]
Associate Professor, Dept. Computer Science, Nehru Arts and Science College, Coimbatore, Tamil Nadu, India. Email: drvimalcs@gmail.com
M.V.Jisha[2]
Ph.D Scholar, Dept. Computer Science, Nehru Arts and Science College, Coimbatore, Tamil Nadu, India. 2Email: jisharudhra@gmail.com

ABSRACT- **Abundant proliferation in usage of credit, debit and ATM card transactions, their use has become increasingly rampant in recent years. The proposed paper work investigates the efficiency of applying classifying algorithms to detect frauds prevailing in its usage. There exists various factors to analyze plentiful classification algorithms ,like KNN, Logistic Regression, Support Vector Machine and Random Forest. The proposed work analyzed that the performance of Random Forest is the efficient algorithm to detect the fraud transaction in terms of different factors .**

Keywords : Credit cards, KNN, Logistic Regression, Support Vector Machine and Random Forest

I. INTRODUCTION

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning ,statistics, and database systems. Modern techniques based on Data mining, Machine learning , Sequence Alignment technique,Fuzzy Logic , Genetic Programming, Artificial Intelligence etc., has been introduced for detecting and preventing credit /ATM card ,cheque book type of fraudulent transactions(A.Shen,et.al,2007). Fraud detection is generally viewed as a classification problem in data mining , where the objective is to correctly classify the credit card transactions as fraudulent. In the present scenario , when the term fraud comes into discussion ,credit card follows in the banks and the financial frauds done by the cash card cloning and various fraud clicks. With the increase in credit cards, ATM cards and E-transactions ,fraud has increasing excessively in recent years. Fraud detection includes analyzing of the spending behavior of users or customer order purpose, uncovering or escaping of undesirable behavior. As credit card becomes the most general mode of payment or both online as well as regular purchase ,fraud relate with it are also accelerate(S.Benson Edwin Raj,et al.2011). Fraud is a millions dough business and it is rising every year. Fraud presents significant cost to our financial prudence measure world wide. In this study, we evaluate to advanced data mining approaches, Random Forest and Support Vector Machines ,together with the well known logistic regression, K nearest neighbors, as part of an attempt to better detect credit card fraud .The study used the dataset from large data transactions.

Statistical fraud detection methods have been divided into two broad categories: Supervised

and unsupervised. In supervised fraud detection methods , models are estimated based on the samples of fraudulent and legitimate transactions, to classify new transaction as fraudulent or legitimate. In unsupervised fraud detection , outliers or unusual transactions are identified as potential cases of fraudulent transactions. Both these fraud detection methods predict the probability of fraud in any given transaction.

The remainder of this paper is organized as follows. Section II gives a brief description on credit card fraud ,the next section describes the methodology. Section IV gives the comparison of classification algorithms using various algorithms, followed by conclusion and references.

II. CREDIT CARD FRAUD

Credit card fraud has been divided into two types: Offline fraud and On-line fraud. Offline fraud is committed by using a stolen physical card at call center or any other place (D.W.Hosmer,et al.,2000). On-line fraud is committed via internet, phone ,shopping ,web, or in absence of card holder.Authorized users are permitted for credit card transactions by using the parameters such as credit card number, signatures, card holders address, expiry date etc . The unlawful use of card or card information without the knowledge of the owner itself and thus in an act of criminal deception refers to credit card fraud. Credit card fraud detection is quite confidential and is not much disclosed publicly as in Fig.1 . Commonly used detection methods are Rule-based techniques, Random Forest, Decision trees, Support Vector machines ,Logistic regression ,ANNs and meta-heuristics such as k-means clustering , Genetic algorithms and nearest neighbors algorithms. Fraud is some kind of human behavior that relate to larceny, misinterpretation, misrepresent, unethical , craftiness false suggestions etc. Many companies deals with millions of external parties, it is cost prohibitive to check the majority of the external parties activity and identity manually. Certainly, for investigating each suspicious transaction, they incur a direct overload cost for each of them. If in case, transaction amount is smaller than overhead cost, investigating is not worthwhile. Transaction involve among banking institutions offering financial transaction services, logistics companies offering various kind of transportation services. These transactions contain sensitive information in the form of data so there must be a technique which is applied on these financial transactions.

The basic goals of information security,

- Privacy : Information must be secret from unauthorized parties.
- Integrity : Assurance that received data not contains any kind of alteration, addition ,scoring through or replay.
- Authentication :The assurance that the communication entity is one that is claimed to be.
- Non-repudiation : Sender and receiver prove their identities to each other.
- Access control : This service controls that have access to a resource and under what condition access can occurs. The frequent use of plastic cash card is the most sensitive and vulnerable part of transaction system. It leads to the violation of all the above security issues and attracts skimmers. But the most important and prominent part is that when the customer access the ATM machine for transaction.

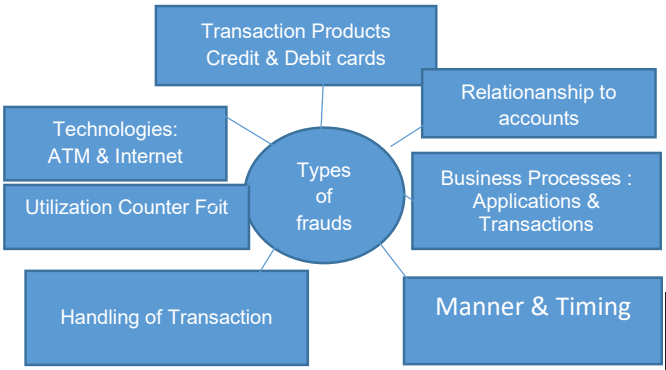

Fig.1. Types of Frauds

III. METHODOLOGY

The methodology describes the different data mining techniques used, the dataset,the tool used and the analytical measures used to evaluate the performance of the mentioned classification algorithms.

A. DATAMINING TECHNIQUES AND THE CONFUSION MATRIX

We investigated the performance of four techniques in predicting fraud: Logistic regression (LR), Support Vector Machines (SVM), Random Forest (RF) and K nearest neighbours (KNN).In the paragraph below, we briefly describe the four techniques employed in this study(Siddhartha Bhattacharya, et al,2010).

*1.) Logistic Regression:*Qualitative response models are appropriate when dependent is categorical. In this study, our dependent variable fraud is binary and logistic regression is widely used in such problems(D.W. Hosmer et al.2000).For example, used binary choice models in the case of insurance frauds to predict the likelihood of a claim being fraudulent . Prior work in related areas has estimated logit models of fraudulent claims insurance, food stamp programs and so forth (V. Vapnik,1998).

The confusion matrix for Logistic Regression algorithm is given in Fig.2 ,which represents the actual detection of 100 % fraud from the dataset.We have the True Positive Value as 21,True Negative value as 633,False Positive value as 51 and False Negative as 557.

Actual class	Predicted class		Row total
	Yes	No	
Yes	557	21	578
No	51	633	684
Column total	608	654	1262

Fig.2. **CONFUSION MATRIX FOR LR**

*2.) Support Vector Machines:*Support Vector Machines are statistical learning techniques that have been found to be very successful in a variety of classification tasks. Several unique features of these algorithms make them suitable for binary classification programs like fraud detection SVMs are linear classifiers that works in a high dimensional feature space that is a non-linear mapping of the input space of the problem at hand. This simplicity of linear classifiers and the capability to work in a feature-rich space make SVMs attractive for fraud detection task where highly unbalanced nature of data (fraud and non-fraud cases) make extraction of meaningful features critical to the detection of fraudulent transaction is difficult to achieve . Applications of SVMs include informatics, machine vision, text categorization and time series analysis .

The confusion matrix for Support Vector Machine algorithm is given in Fig.3 ,which represents the actual detection of 100 % fraud from the dataset.We have the True Positive Value as 41,True Negative value as 634,False Positive value as 20 and False Negative as 567.

Actual class	Predicted class		Row total
	Yes	No	
Yes	567	41	608
No	20	634	654
Column total	587	675	1262

Fig.3. **CONFUSION MATRIX FOR SVM**

*3.) K-nearest neighbor:*K-nearest neighbor algorithm is a non-parametric method used for classification and regression .In both case ,the input consist of the K closest training examples in the feature space. The output depends on whether K-NN is used for classification or regression.Both for classification and regression, a useful technique can be used to assign weight to the contribution of the neighbor, so that the closest neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consist in giving each neighbor a weight of 1/d where d is the distance to the neighbor. The neighbors are taken from a set of objects for which the class(for k-NN classification) or the object property value (k-NN regression) is known. This can be thought of as training set for the algorithm, though no explicit training step is required. It is used in application for face detection mean shift tracking analysis and typical computer vision problems.

The confusion matrix for K-nearest neighbours algorithm is given in Fig.4 ,which represents the actual detection of 100 % fraud from the dataset.We have the True Positive Value as 209,True Negative value as 445,False Positive value as 207and False Negative as 401.

Actual class	Predicted class		Row total
	Yes	No	
Yes	401	209	610
No	207	445	652
Column total	608	654	1262

Fig.3. **CONFUSION MATRIX FOR KNN**

*4.) Random Forests:*The popularity of decision tree models in data mining arises from their ease of use, flexibility interms of handling various data attribute types, and interpretability . Single tree models , however, can be unstable and overly sensitive to specific data . Ensemble methods seek to address this problems by developing a set of models and aggregating their predictions in determining the class label for a data point . A Random Forest(Raghavendra Patidar et,al.,2011) model is an ensemble of classification (or regression) trees. Ensembles perform well when individual members are dissimilar , and Random Forests obtain variation among individual tress using two sources for randomness : first , each tree is build on bootstrapped samples of the training data; secondly, only a randomly selected subset of data attributes is considered at each node in building the individual trees. Random Forest thus combine the concept of bagging , where individual models in an ensemble are developed through sampling with replacement from the training data, and the random subspace method, where each tree in an ensemble is build from a random subset of attributes. Random Forest are computationally efficient since each tree is build independently of the others . Theyhave been applied in recent years across varied domains from crediting customer churn(Jisha.m.v et al,2018), image classification, to various bio-medical problems.

The confusion matrix for Random Forest algorithm is given in Fig.4 ,which represents the actual detection of 100 % fraud from the dataset.We have the True Positive Value as 4,True Negative value as 650,False Positive value as 37 and False Negative as 571.

Actual class	Predicted class	Row total	
	Yes	No	
Yes	571	4	575
No	37	650	687
Column total	608	654	1262

Fig. 4. **CONFUSION MATRIX FOR RANDOM FOREST**

B. Dataset

A Comparison of the classification algorithms – Random Forest , KNN, Support Vector Machine and Logistic Regression by using the dataset containing 31 attributes and 3 lakhs (approx.) transactions for detecting the frauds .Analysis is done by highlighting the 16 factors ,to detect the best fraud detecting classification algorithm using the R tool(Jisha.M.V,2018).

C. Tools Used

R studio software is used in our work for the analysis of various algorithms. In Rsudio it is very easy to install required packages because of its user friendly behavior. It is an open source integrated development environment (IDE) for R programming language. R language provides a wide variety of statistical and modern graph techniques. It is very easy to understand and implanting a code with this tool.

D. Analytical measures

The performance of the proposed system was evaluated using Precision, Recall, F-Measure and Kappa Statistics. Precision, Recall and F-Measure were calculated using the result of a confusion matrix.

- Precision = TP/(TP=FP)
- Recall = TP/(TP=FN)
- F-Measure = 2 *[(Precision * Recall) / (Precision + Recall)]
- Sensitivity= TP/FP
- Specificity=TN/FN

Along with the above factors ,the other evaluated factors are Accuracy, Kappa, Prevalence ,detection rate, detection prevalence, P -value, Positive Pred value ,Negative Pred value , Mcnemar's Test P-Value, AUC etc. The above mentioned factors are used for evaluating the performance of the four classifiers.

*1) Comparison of the classifiers using precision:*Precision measures the number of true positives divided by the number of true positives and false positives. In other words, precision is the measure of a classifier exactness. Table 3 presents the precision values of the four classifiers. It was observed that the Random forest has higher precision values than the other classifiers. A lower precision indicates large number of false positives. It is therefore inferred that other classifiers has more non-fraudulent transactions labeled as fraudulent.

*2) Comparison of the classifiers using recall:*Recall measures the number of the true positives divided by the number of true positives and the number of false negatives. In essence, recall can be thought of as a measure of the classifier completeness. A low recall indicates many false negatives(Oluwafolake Ayano, et al,2017).

*3) Comparison of the classifiers using F-Measure:*The F-Measure indicates the balance between the recall and precision values.

*4) Comparison of the classifiers using Kappa statistics:*Kappa statistics represent the extent to which the data collected correctly represents the variables measured.

*5) Comparison of the classifiers using Sensitivity and Specificity:*Sensitivity compares the amount of items correctly identified as fraud to the amount incorrectly listed as fraud ,also known as the ratio of true positives to false positives. Specificity refers the same concept with legitimate transactions, or the comparison of true negatives to false negatives.

6) Comparison of the classifiers using P-Value: A small p-value (typically <=0.05) indicates strong evidence against the null hypothesis, so we reject the null hypothesis. A large p-value (>0.05)indicates weak evidence against the null hypothesis, so we fail to reject the null hypothesis.

7) Comparison of the classifiers using Mc Nemar's test: It is a statistical test used on paired normal data. It is applied to 2*2 contingency tables with a dichotomous trait, with matched pairs of subjects, to determine whether the row and column marginal frequencies are equal. It is named after Quinn McNemar, who introduced it in 1947.

*8) Comparison of the classifiers using Prevalence:*Prevalence is a term which means being widespread and it is distinct from incidence. Prevalence is a measurement of all individuals affected by the disease at a particular time, whereas incidence is a measurement of the number of new individuals who contract a disease during a particular period of time.

9) Comparison of the classifiers using Area under the curve: AUC is the abbreviation for area under the curve. It is used in classification analysis in order to determine which of the used models predicts the classes best. An example of its application are ROC curves. Here, the True positive rates are plotted against False positive rates.

10) Comparison of the classifiers using Detection rate: Detection rate is mainly reflected in confusion matrix. It is a parameter that will vary according to the dataset. Detection rate =TP/(TP+FP+FN+TN).

11) Confusion Matrix Representation for Classification Algorithms: Confusion Matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. This matrix itself is relatively simple to understand , but the related terminology can be confusing.Each confusion matrix for each classification algorithm is given above.

IV. Comparision Of Classification Algorithms Using Various Factors

The Dataset is divided ,to detect the fraud from the large data transactions.Table1 shows the performance of the algorithms on detecting the 25 % fraud, we could analyse that for a small data set ,SVM is having high accuracy ,followed by Random Forest ,Logistic Regression and KNN.

Table 1
Dataset With 25 % Fraud Data

FACTORS	KNN	LOGISTIC REGRESSION	RANDOM FOREST	SVM
Accuracy	0.6518	0.9393	0.9521	0.9617
95% CI	0.5961, 0.7045	0.9068, 0.9631	0.9222, 0.9729	0.934, 0.98
No Information Rate	0.5335	0.5335	0.5335	0.5463
P-Value [Acc > NIR]	1.49e-05	< 2e-16	< 2.2e-16	<2e-16
Kappa	0.3037	0.8777	0.9033	0.9228
Mcnemar's Test P-Value	0.3382	0.1687	0.009823	0.3865
Sensitivity	0.6644	0.9110	0.9110	0.9718
Specificity	0.6407	0.9641	0.9880	0.9532

Pos Pred Value	0.6178	0.9568	0.9852	0.9452
Neg Pred Value	0.6859	0.9253	0.9270	0.9760
Prevalence	0.4665	0.4665	0.4665	0.4537
Detection Rate	0.3099	0.4249	0.4249	0.4409
Detection Prevalence	0.5016	0.4441	0.4313	0.4665
Balanced Accuracy	0.6526	0.9375	0.9495	0.9625
'Positive' Class	Yes	Yes	Yes	Yes
AUC	0.652	0.941	0.956	0.963
Precision	0.5512	0.9446	0.9547	0.9383
F-measure	0.6025	0.9274	0.9323	0.9547

Table 2 gives the comparison of classification algorithms for detecting 50% fraud from the given dataset.It is viewed that Random Forest is having high accuracy compared to other algorithms.From this point ,We could view that as the size of the dataset increases, the performance of SVM decreases.

Table.2
Dataset With 50 % Fraud Data

FACTORS	KNN	SVM	LOGISTIC REGRESSION	RANDOM FOREST
Accuracy	0.6248	0.9686	0.9686	0.9733
95% CI	0.5859, 0.6625	0.9519, 0.9807	0.9519, 0.9807	0.9576, 0.9844
No Information Rate	0.5259	0.529	<2e-16	0.5259
P-Value [Acc > NIR]	2.995e-07	<2e-16	0.9371	<2e-16
Kappa	0.2477	0.937	0.5259	0.9464
Mcnemar's Test P-Value	0.6917	0.8231	0.5023	0.332
Sensitivity	0.6060	0.9700	0.9735	0.9636
Specificity	0.6418	0.9674	0.9642	0.9821
Pos Pred Value	0.6040	0.9636	0.9608	0.9798
Neg Pred Value	0.6437	0.9731	0.9758	0.9676
Prevalence	0.4741	0.4710	0.4741	0.4741
Detection Rate	0.2873	0.4568	0.4615	0.4568
Detection Prevalence	0.4757	0.4741	0.4804	0.4662
Balanced Accuracy	0.6239	0.9687	0.9688	0.9728
'Positive' Class	Yes	Yes	Yes	Yes
AUC	0.624	0.969	0.968	0.974
Precision	0.6545	0.9513	0.9729	0.9965
F-Measure	0.6292	0.9604	0.9731	0.9797

Table 3 gives the comparison of algorithms with 75% fraud data from the dataset .It is viewed that that Random Forest is having high accuracy of 0.9776 ,followed by SVM with 0.97.There is a variable difference between the two algorithms .But comparing other factors, Random Forest algorithm is good in its performance than the other algorithms.

Table 3
Dataset With 75 % Fraud Data

FACTORS	KNN	SVM	LOGISTIC REGRESSION	RANDOM FOREST
Accuracy	0.6457	0.9701	0.9594	0.9776
95% CI	0.6141, 0.6763	0.9571, 0.9801	0.9448, 0.9711	0.9659, 0.9861
No Information Rate	0.5197	0.5304	0.5197	0.5197
P-Value [Acc > NIR]	4.604e-15	< 2e-16	< 2e-16	< 2e-16
Kappa	0.2888	0.9401	0.9187	0.955
Mcnemar's Test P-Value	0.2068	0.08897	0.01496	0.00225
Sensitivity	0.6044	0.9795	0.9400	0.9600
Specificity	0.6838	0.9618	0.9774	0.9938
Pos Pred Value	0.6385	0.9578	0.9747	0.9931
Neg Pred Value	0.6517	0.9815	0.9463	0.9641
Prevalence	0.4803	0.4696	0.4803	0.4803
Detection Rate	0.2903	0.4600	0.4514	0.4610
Detection Prevalence	0.4546	0.4803	0.4632	0.4642
Balanced Accuracy	0.6441	0.9707	0.9587	0.9769
'Positive' Class	Yes	Yes	Yes	Yes
AUC	0.645	0.971	0.960	0.979
Precision	0.6312	0.9088	0.9810	0.9952
F-Measure	0.6173	0.9427	0.9599	0.9771

Table 4 gives the comparison of the algorithm for detecting 100% fraud data from the large dataset.It is viewed that that Random Forest is having high accuracy of 0.9675 ,followed by SVM with 0.9517.Finally considering all other factors, Random Forest algorithm is good in its performance than the other algorithms in detecting fraud from a large data transactions

Table 4
Dataset With 100 % Fraud Data

FACTORS	KNN	SVM	LOGISTIC REGRESSION	RANDOM FOREST
Accuracy	0.6704	0.9517	0.9429	0.9675
95% CI	0.6437, 0.6963	0.9383, 0.9628	0.9287, 0.9551	0.9562, 0.9766
No Information Rate		0.5349	0.5182	0.5182
P-Value [Acc > NIR]	<2e-16	< 2e-16	< 2.2e-16	< 2.2e-16
Kappa	0.3399	0.9031	0.8855	0.9348
Mcnemar's Test P-Value	0.9609	0.01045	0.0006316	5.806e-07
Sensitivity	0.6595	0.9659	0.9161	0.9391
Specificity	0.6804	0.9393	0.9679	0.9939
Pos Pred Value	0.6574	0.9326	0.9637	0.9930
Neg Pred Value	0.6825	0.9694	0.9254	0.9461
Prevalence	0.4818	0.4651	0.4818	0.4818
Detection Rate	0.3177	0.4493	0.4414	0.4525
Detection Prevalence	0.4834	0.4818	0.4580	0.4556
Balanced Accuracy	0.6700	0.9526	0.9420	0.9665
'Positive' Class	Yes	Yes	Yes	Yes
AUC	0.670	0.953	0.945	0.970
Precision	0.6672	0.9523	0.9643	0.9915
F-measure	0.6633	0.9590	0.9394	0.9645

IV. CONCLUSION

In this paper , we have brief description discussion on the credit card fruad detection using four classifiers. The classifiers detects the fraud, from a large data transactions , using various factors The confusion matrix shows that among the classifiers, the Random Forest algorithm is efficient with high accuracy of 0.9675. The relative studies and our results arrive to the fact of detecting a proposed and efficient algorithm than random forest which is our next proposed work.

ACKNOWLEDGEMENT

We Authors would like to thank all the contributors of the journal.All the authors whose articles helped us for our work, for reference , listed in reference . Wish our paper will be a reference for future scholars.

REFERENCES

[1]. Srivastava, Aman, et al , "Credit card fraud detection at merchant side using neural networks." Computing for sustainable global development (INDIACom),2016 3rd International Conference on ,IEEE, 2016.

[2]. Dal Pozzolo , Andrea, et al. "Credit card fraud detection and concept-drift adaptation with delayed supervised information ."Neural Networks (IJCNN),2015 International Joint Conference on, IEEE,2015.

[3].A.Shen, R.Tong, and Y.Deng, "Application of classification models on redit card card fraud detection",June 2007.

[4].Abhinav Srivastava, Amlan Kundu, Shamik Sural and Arun K. Majumdar, "Creditcard Fraud Detection Using Hidden Markov model"IEEE, Transactions on Dependable and Secure Computing, vol.5, No 1., January-March 2008.

[5]. S.Benson Edwin Raj, A. Annie Portia " Analysis on Credit Card Fraud Detection Methods."IEEE- International Conference on Computer, Communication and Electrical Technology; (2011).(152-156).

[6]. Raghavendra Patidar, Lokesh Sharma, "Credit Card Fraud Detection Using Neural Networks " , International Journal of Soft Computing and Engineering (IJSCE),2011.,Volume -1 Issue ,(32,38).

[7] .Bhattacharya.S., et al. (2011). Data mining for credit card fraud :a comparative study. Decision support systems, vol.50, no.3, pp. 602-613.

[8]. Prajel Save, Pranali Tiwarekar , Ketan N.Jain , Neha Mahyavanshi , " A Novel Idea for credit card fraud detection using decision tree",International Journal of Computer Applications.,volume 161-no13,March-2017.

[9]. Pratiksha .L. Meshram, Parul Bhanarkar, "Credit and ATM Card Fraud Detection Using Genetic Approach",International Journal of Engineering Research & Technology (IJERT),Vol.1 Issue 10, December-2012.

[10].D.W. Hosmer, S. Lemeshow, Applied Logistic Regression, 2nd Ed,Wiley-Interscience,2000.

[11]. V. Vapnik, Statistical Learning Theory, Wiley, New York., 1998.

[12]. MJ Kim and T.S Kim, "A Neutral classifier with Fraud density Map for Effective Credit Card Fraud Detection", Proc Int "l Conf.Intelligent Data Eng and Automated Learning. Pp. 378-383,2002.

[13] K. RamaKalyani, D.UmaDevi, "Fraud Deyection of Credit Card Payment System by Genetic Algorithm",International Journal of Scientific & Engineering Research Volume 3, Issue 7, July-2012.

[14] Ekrem Duman, M.Hamdi Ozcelik " Detecting credit card fraud by genetic algorithm and scatter search", Elsvier, Expert Systems with Applications, (2011). 38; (13057-13063)

[15]. Jisha.M.V, Dr.D. Vimal Kumar, " An Efficient Credit Card Fraud Classifier of the four data mining classification algorithms- A Comparative Analysis."(JETIR)Journal of Emerging Technologies and Innovative Research. Nov.2018.

[16]. Siddhartha Bhattacharya, et al, "Data mining for credit card fraud : A comparative study",Decision Support Systems ,(2010).

[17]. Oluwafolake Ayano, et al," A multi-algorithm data mining classification approach for bank fraudulent transactions." Academic journals - African journal of mathematics and comuter science research,2017.