Credit Card Fraud Detection using Isolation Forest, Local Outlier Factor and Support Vector Machine

Kiran Kamari[*1], Dr. Parul Gupta[2], Ms. Monika Gupta[3]

Mtech Student*[1], Assistant Professor[2], Assistant Professor[3]

kiran6652@gmail.com,parulgupta_gem@yahoo.com,monika.mittal167@gmail.com

9650775339, 9810588553, 8950602237

Department of Computer Engineering,

J.C. Bose University of Science and Technology,

YMCA, Faridabad, Haryana

Abstract- Nowadays e-commerce and online transaction is growing rapidly. For online and offline transaction most of the customer uses credit card. Credit card used globally for online transaction, buy goods, product, and payment. The rising use of credit card can increase the chances of fraud in credit card. Credit card system is at risk now. The effect of these fraudulent transaction is on the bank and institute causing a financial loss to them. To detect the fraud in credit card transaction we used different machine learning techniques. The aim of this paper is to identify the fraudulent transaction and outlier in credit card transaction. The dataset of credit card is unbalanced. There are various techniques by which fraudulent transaction can be detected and we have used these techniques such as isolation forest method, local outlier factor and support vector machine to determine fraud in credit card. Used different matrices for enhancing the performance and accuracy. At last comparison analysis is done by using isolation forest, local outlier and support vector machine which give the better result.

Keywords- credit card, Dataset, Fraud detection, isolation forest, Local outlier factor, support vector machine.

I INTRODUCTION

In our daily life the problem of credit card fraud is taking place at a very high rate. The rate of fraudulent activities is rapidly increasing. We can use credit card by both offline and online mode. As now the transaction is online based but then also there is fully chance of credit card fraud. Due

to which people going through a lot of problem. The fraudulent activities causing financial loss to many organization, companies, and government agencies. The credit card fraud can occur by stealing other person credit card or to if someone giving the number of credit card to other. The numbers can be increased in future more so to prevent this many researchers focusing in this field. They giving their approach to detect the fraudster. The credit card fraud is occurring due to following two reasons. First is that the behaviors of fraudulent attempting and second is the highly imbalanced data. To tackle this problem many techniques and approaches are given.

The dataset which is used here are taken from kaggle which consist transaction done by credit card by the customer in September 2013 in Europe. Dataset is highly imbalanced Credit card transaction is categorizing in two category i.e. fraudulent and non-fraudulent. These two classes create anomalies. Which can be detect by using machine learning algorithm. So in this paper we are proposing a various or different machine learning techniques to detect fraudulent transaction in credit card. The different techniques that we are applying are isolation forest, local outlier factor and support vector machine. Which used to find the outlier in credit card transaction. For classification task we used different matrices.

Rest of the paper is systematized in five sections. Sections II describe the existing literature. Section III explain the proposed architecture. Section IV explain the detail of dataset. Section V describe the detail explanation of methodology. Section VI show the experimental setup and Results. Section VII show the conclusion of this paper.

II LITERATURE REVIEW

Changjun Jiang [1] to see the transaction behavior of cardholder it proposed a novel fraud detection method in which it uses the historical transactions data of cardholder and divide them into various groups. It find that the behavior of same groups of member are similar. on the basis of transactions amount it distribute cardholder into three levels, low, medium and high with the help of clustering method.in each of group the transaction is combined with the help of proposing a theory of window sliding. The cardholders behavior form is characterize by deriving surplus feature taken from window. On the basis of combine transactions for every single cardholder it remove assemble behavioral patterns and then classifier are trained for every group respectively. Detection of fraud online takes place by classifier and if it found a forge in a newly transactions then feedback mechanism plays role to resolve the drift problem.

Eunji Kima, Jehyuk Lee [2] proposed a hybrid ensemble method. To handle the card and association issue it conclude the deep learning method.it introduced two framework, champion and challenger. Respectively both champion and challenger is the attitude of hybrid ensemble and deep learning model. To estimate the large amount of real world transactions dataset it manipulated several realistic metrics. For the post-lunch execution it set up these models into FDS after completing the offline testing. The winning model divided into two parts one is off-line and other one is post-launch testing. The main task of champion and challenger framework task is to compare the two models.it was also used deep learning model to work with FDS.

Vaishnavi Nath Dornadula [3] propose a method called a fraud detection method used for streaming transactions data. The objective of this is to examine the details of previous transactions and the behavioral pattern is to remove of the customer. For detecting fraud multiple supervised and semi supervised techniques are use. Based on transactions amount cardholder are clustered in various groups. Sliding window technique is used to remove the behavioral pattern of the groups. Next is to train the classifier. And then that classifier will be chosen which has better rating to detect fraud. Generally in real world researcher resolve the concept of drift problem, to deal with the imbalance data it used Matthews's correlation coefficient. And to balance the dataset it used SMOTE, where the classifier were working better than in case of previous one. To deal with imbalance data it can also use one class classifier. For better performance and outcome it use random forest, decision tree, and logistic regression.

Ankit Mishra [4] focus to misclassifying transactions. Nowadays all the payment and transfer of money is done by credit card due to this the possibility of fraud is increases. Cybercriminal and hackers done the fraud in transction.so the author try to judge the several classifier and metric by examine the different classification problem. The model deal with the problem of genuine transaction which is fraud and it also find the fraud.

Yvan Lucas, P.E. Portier, L. Laporte [5] For the problem of sequence classification we proposed the sate-of-the art methods.it uses HMM technique.to increase the effectiveness of classification task it used a HMM feature model.to check whether transactions is fraudulent or not in credit card it acquire the state of the art approach. Presenting the experimental result and ecommerce transaction with various classifier. It demonstrate the heftiness of the approach by a hyper parameter.by relating various solution the issue of structural missing values can be resolved. HMM

based feature is used to detect the anomaly in credit card transactions. This work is helpful for feature engineering in sequential data.

III PROPOSED TECHNIQUE

The proposed technique are used to identify the frauds in credit card transcation.in this paper we detect the fraudulent transaction and genuine transaction. Apply different machine learning techniques such as Isolation forest, local outlier factor and support vector machine. Comparison analysis is doing by using these machine learning techniques and finding the best technique which give the highest accuracy and good performance Figure 1 show the flow work of system architecture.

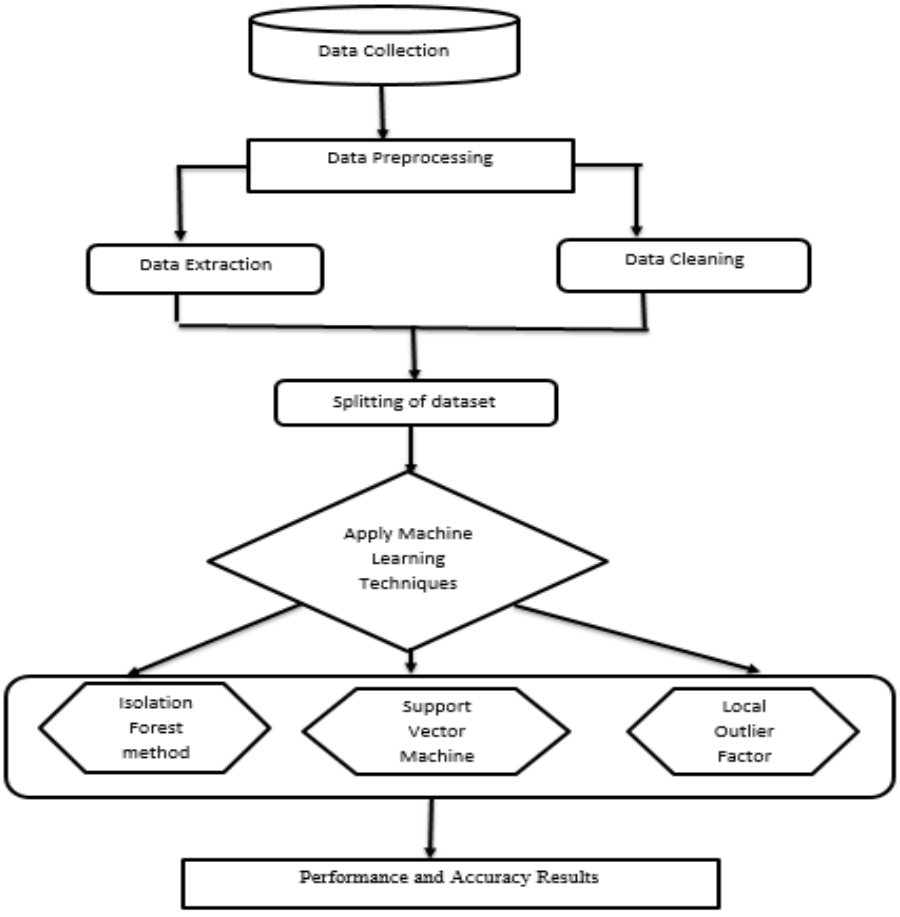

Figure 1 system architecture

IV DATASET

The dataset which we used in our research is founded in kaggle. The dataset consist of transaction done by European cardholders through credit cards in September 2013. The dataset brings out the transaction that is arisen in two days where we found 492 fraud from 284,807 transaction. This dataset is largely imbalanced where the fraud account for 0.172 percent transaction. The result of PCA transformation is input numeric variables in dataset. Because of confidentially matters it does not give original feature and information of background of the data. The principle components such as feature V1, V2, V28 are achieved with PCA. Time and Amount are the only feature which cannot be transformed with PCA. The feature time can be defined as consisting of the second which elapsed in between every transaction respectively in the dataset. Whereas the feature amount can be defined as the transaction amount which helps in dependent cost learning. Class is defined to give response to variable now if it gives value 1 it means the case is fraud and otherwise it will give value 0.

V METHODOLOGY

A. Isolation Forest Method

Isolation method focuses on something unusual and different, so which is why it is called anomaly and we use it here for recognition of unusual patterns. So eventually we call it Anomaly Detection algorithm. These days, Anomaly Detection is having wide application across all industries (Banking, Finance, Healthcare, Manufacturing and Networking). This works same as Decision Tree algorithm, which starts with node root and keep going on other space. Say suppose, we have to identify a mole in something similar data set by eyes, we will not be able to identify. Right? Therefore, I have used such methods to find and improve the fraud detection in credit card use. Isolation forest.

If in a large dataset many data are same and one of them is different from others, this method helps in isolating anomalous data in whole dataset. And main benefits of this method is chances of exploiting sampling methods to a dimension which are not allowed to profile based methods. Isolation Forest method aims on anomalies data as it has kind of shorter path as compared to profiling normal data. It mainly helps in creation of fast algorithm to detect the anomalies with very less in memory consumption.

Algorithm is given to figure out the anomalous part.

- Create a profile those data which is normal
- Observe the whole csv data clearly import it
- Report anything which can't be retained as normal
- Use the formula to get the anomalies score

$$s(x,n) = 2^{-E(h(x))/c(n)} \text{And } c(n) = 2H(n-1) - 2(n-1)/n)$$

- Where n is number of the data points and c (n) is the average path length of unsuccessful search in a Binary search tree. It normalizes the score in between 0 to 1.
- when $E\big(h(x)\big) \to c(n), s \to 0.5$;
- when $E\big(h(x)\big) \to 0, s \to 1$;
- and when $E\big(h(x)\big) \to n-1, s \to 0$;
- Identify the score. So, if score is closer to 1, then it is an anomalous point otherwise it's a normal data.

So, once we find the anomalous part, it isolates the anomalies explicitly in a large data set. Isolation Forest method creates multiple partitions on whole dataset based on random selected patterns and randomly split those by those patterns and features. It normalizes the dataset in different small partitions. And it has become less lousy to find out the final output.

B. Local Outlier Factor Method

An Outlier is the other data point that differs profusely from the rest of the existing observations in a dataset. Local outlier factor method comes in mind where data is multidimensional. It basically helps in detecting the outliers in data. This method is based on density which relies on closest neighbor search. Local Outlier Factor (LOF) is also having an algorithm for anomalies detection which calculates the local density deviation for given data with comparison of its neighbors. Outliers are some patterns which do not come as expected outcome. So, figuring out such patterns are the main important objective here in credit card fraud detection.

Outlier detection mainly works in data analysis, anomalies finding, and also helps out to discover upcoming activities in the crucial safety systems. It helps in pre-prediction of many fraudulent activities like credit/debit card theft, claiming fake insurance, stealing taxes, monitoring real time systems, medical areas and various online transactions. So here used Local outlier factor method especially for credit cards fraud detection. So, I have used K-nearest neighbor's detector and followed some algorithm steps to find the local outlier.

In general, I am taking n-neighbor = 20 by default to work correctly. And I have used 'auto' algorithm for calculating n-neighbor and the metric and auto algorithm is consisting of few steps which are given below.

- Calculate distance between points
- Calculate Kth-Nearest Neighbor Distance
- Calculation of K-Nearest Neighbor
- Calculation of Local Reachability Density (LRD)
- Calculation of Local Outlier Factor
- Finally, Results Analysis

So, if n-neighbor is having value of 20 or greater than 20 that is normal observation and if it comes less than 20, then that will be counted as outliers.

C. Support Vector Machines

In the field of Machine Learning, Support Vector machine is the most available algorithm which is used for regression and classification. But it is mainly used for classification-based problems in other aspects of fraudulent field. It has a very different implementation approach as compared to other algorithms which are present in Machine learning and it has capacity to manage various category's variables simultaneously. The purpose of Support Vector Machine algorithm is to build the best decision boundary which can help in n-dimensional space segregation into classes, so that we can easily get the fresh data points in right categories in future and here most famous category is known as hyper plane.

In other words, we can say that, it basically helps in choosing the best vectors or points in space for creation of hyperplane and those best vectors are known as Support Vectors. So, that is why this algorithm is known as Support Vector Machine algorithm.

Margin: it shows the gap between those two classes A and B which is separated by line.

Hyperplane: The line which separates the two sets of objects in different categories (Class A and Class B).

Support vectors: Those data points which are nearest to hyperplane are called as Support vectors. So, if we talk about types of Support Vector Machine, we will get two types which are linear and Non-Linear. I have already given an example of Linear. If it is possible to separate two classes with a straight line that means it's a Linear and it's not possible to divide with straight line that is called non-linear. So, in term of solution, I have used kernel trick. Kernel uses the fraud transactional data set to create the classification model and it use the Support vector machine to find out the fraud data.

VI EXPERIMENTAL SETUP AND RESULTS

Evaluation Metric

There are many classification task use evaluation metric. Evaluation metric are used for the purpose of accuracy and performance of system. For better result and improving the performance used metric. To finding the fraud and genuine transaction we need some standard measurement tools are Precision, Recall, F1-score, and Support and confusion matrix.

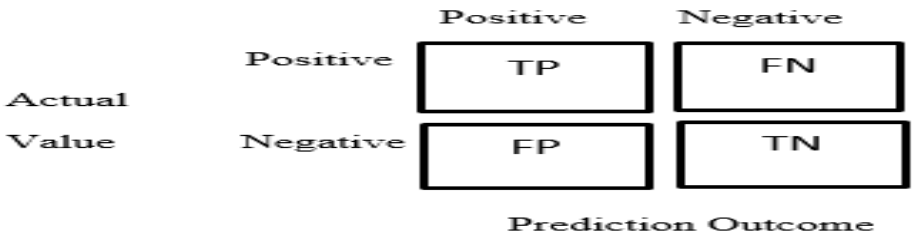

Figure 2 Confusion matrix

- **True Positive (TP):** TP is the value where all the value of predictive and actual value is positive i.e. both actual and predictive value is positive.
- **False Negative (FN):** FN is a value where actual value is negative and predictive value is positive.

- **True Negative (TN):** TN is a value where both actual and predictive value is negative.
- **False Negative (FN):** FN is a value where actual value is positive and predictive value is negative.

Recall: Recall is the ratio of true positive to the actual positive situation which is show in equation i.e. recall is a value of true positive which find in out of all positive situation. It is also called sensitivity

$$\text{Recall} = \frac{\text{TP}}{TP + FN}$$

Precision: It is the ratio of true positive over the true positive and false positive which is shown in equation i.e. how many found cases are true positive.

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

F_1 Score: F1 score is also known as F-measure is the harmonic mean of recall and precision. Its value in between 0 and 1. Where 1 show best and 0 show worst. Shown in equation

$$\text{F1} = \frac{2 * (\text{Precision} * \text{Recall})}{Precision + Recall}$$

Support: it is the number of rate of each class in correct object values.

Table 1 Classification report of isolation forest

	Precision	Recall	F1-score	Support
0	1.00	1.00	1.00	28432
1	0.22	0.22	0.22	49

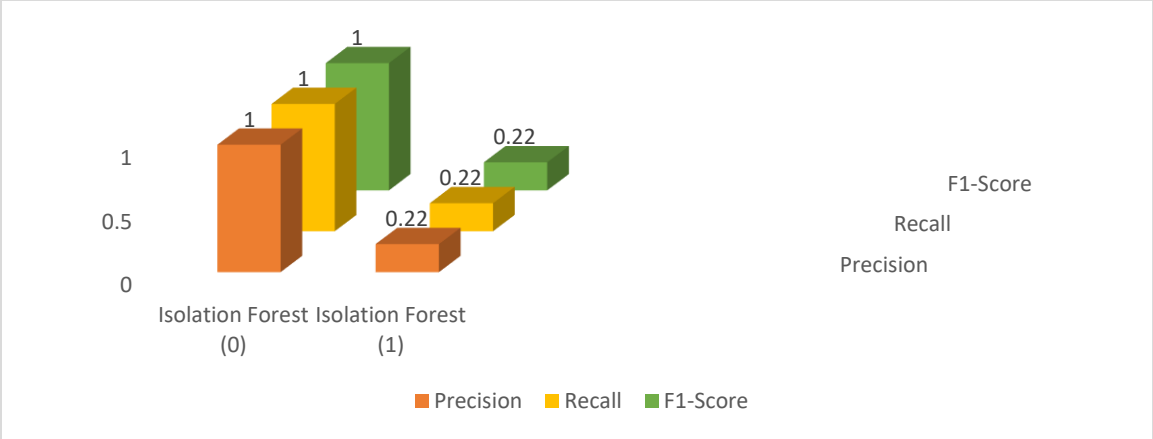

Fig 2: Representation of precision, recall, and f1-score in isolation forest

Table 2 Classification report of Local Outlier Factor

	Precision	Recall	F1-score	Support
0	1.00	1.00	1.00	28432
1	0.02	0.02	0.02	49

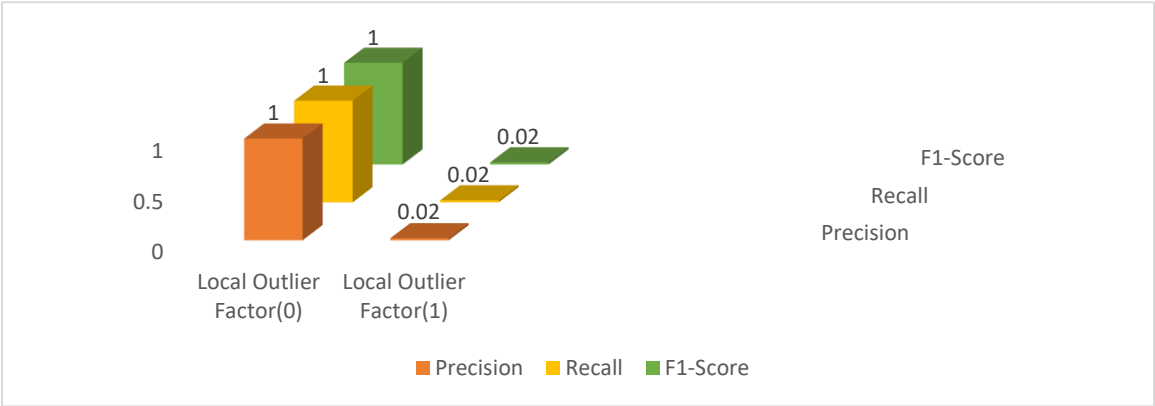

Fig 3: Representation of precision, recall, and f1-score in local outlier factor

Table 3 Classification report of Support Vector Machine

	Precision	Recall	F1-score	Support
0	1.00	0.70	0.82	28432
1	0.00	0.37	0.00	49

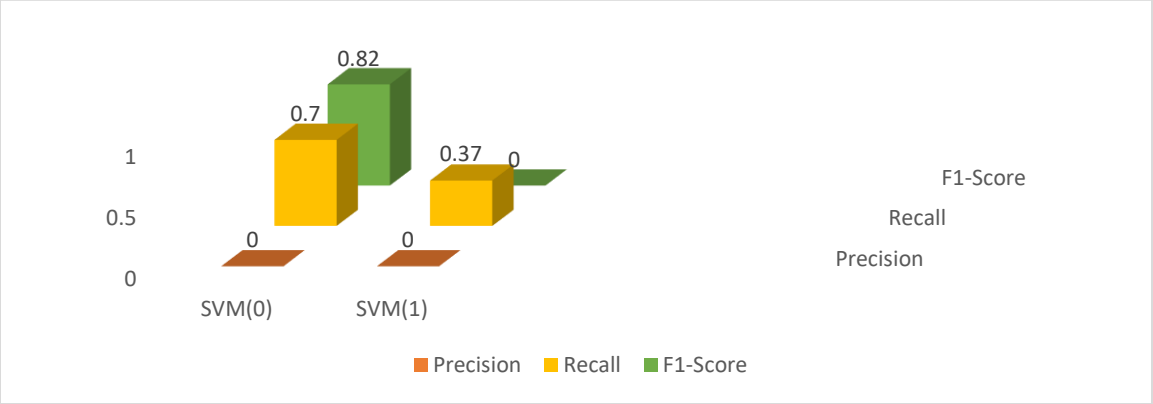

Fig4: Representation of precision, recall, and f1-score in support vector machine

B. Experimental Results.

By comparing all the three models in this graph on the bases of accuracy which showing that Isolation forest method is the best method to detect fraud in credit card transaction. Isolation forest techniques giving the highest accuracy i.e. 99.74% as compared to local outlier factor and support vector machine.

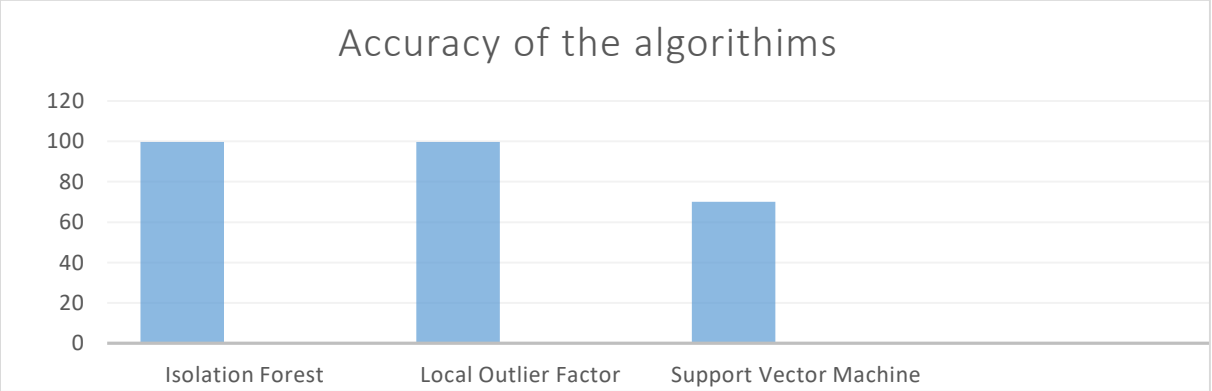

Figure 5 Graph representing accuracy of various fraud detecting algorithms

In table 4 it shows the accuracy of all three techniques such as isolation forest, support vector machine and local outlier factor.

Table 4: Comparison of various techniques

Techniques	Accuracy
Isolation Forest	99.74
Local Outlier Factor	99.65
Support Vector Machine	70.09

This pie chart shows the percentage of fraud detection method by using different machine learning techniques. Local outlier factor shows 37 %, fraud detection rate, support vector machine shows 26% and isolation forest method also show the fraud detection rate.

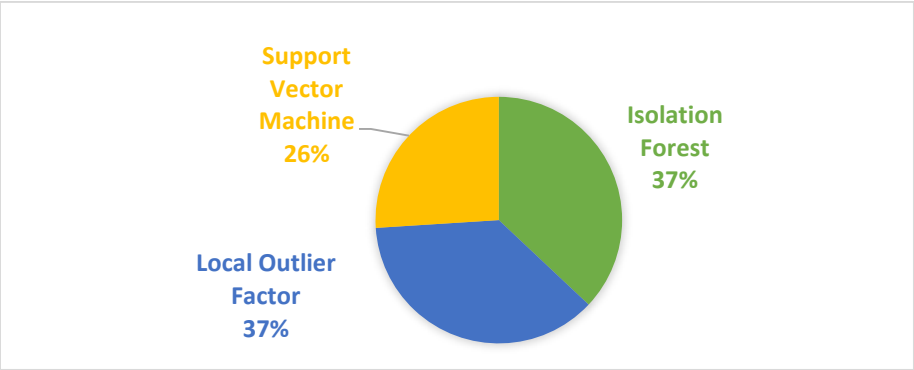

Figure 6 Fraud detection rate

VII CONSLUSION AND FUTURE SCOPE

The paper proposes three different machine learning techniques that focus on outlier detection. The technique which used in this paper are isolation forest, local outlier factor and support vector machine. The paper examined the performance of credit card transaction on the bases of these techniques. Isolation forest give the highest accuracy as compared to local outlier factor and support vector machine. The accuracy of isolation is 99.74%, local outlier 99.65% and SVM 70.09%. On the bases of the performance and accuracy isolation forest is the best method to detect the fraud in credit card transaction. Future work can be done in deep learning in term of increasing accuracy by enhancing the sample size, at the cost of computational expense.

REFERENCES

[1] Tanupriya Choudhary," An Efficient Way to Detect Credit Card Fraud Using Machine Learning Methodologies" IEEE Second International Conference on Green Computing and Internet of Things, pp. 591-597, 2018.

[2] Fabrizio Carcillo," Combining unsupervised and supervised learning in credit card fraud detection" Elsevier Inc. pp.1-15, May 2019.

[3] Sahil Dhankhad, "Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study" International Conference on Information Reuse and Integration for Data Science, IEEE, pp.122-125, 2018.

[4] N. Malini, "Analysis on Credit Card Fraud Identification Techniques Based On KNN and Outlier Detection" 3rd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio- informatics, IEEE, 2017.

[5] Maja Puh, "detecting Credit Card Fraud using Selected Machine Learning Algorithms" MIPRO, pp. 1250-1255, May 2019.

[6] Sangeeta Mittal, "Performance Evaluation of Machine Learning Algorithms for Credit Card Fraud Detection" 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp.320-324, 2019.

[7] M. Suresh Kumar," Credit Card Fraud Detection Using Random Forest Algorithm" 3rd International Conference on Computing and Communication Technologies (ICCCT), IEEE, pp.149-153, 2019.

[8] Changjun Jiang, Jiahui Song, "Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism" IEEE Internet of Things Journal, March 2018.

[9] Eunji Kima, Jehyuk Lee," Champion-challenger analysis for credit card fraud detection: Hybrid ensemble and deep learning" Elsevier Ltd, pp.214-224, 2019.

[10] Vaishnavi Nath Dornadula," Credit Card Fraud Detection using Machine Learning Algorithms " International conference on recent trends in advanced computing (ICRTAC) ,pp. 631–641,2019.

[11] Anuruddha Thennakoon, "Real-time Credit Card Fraud Detection Using Machine Learning" 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE, pp.488-493, 2019.

[12] Aman Srivastava," Credit Card Fraud Detection at Merchant Side using Neural Networks" 3rd 2016 International Conference on Computing for Sustainable Global Development (INDIACom)" IEEE, pp.667-676, 2016.

[13] Ong Shu Yee," Credit Card Fraud Detection Using Machine Learning as Data Mining Technique" e-ISSN: 2289-8131, Vol. 10, pp.1-4, 2016.

[14] Dejan Varmedja," Credit Card Fraud Detection - Machine Learning methods" IEEE, May 2019.