Credit Card Fraud Detection Using Machine Learning

Sarthak Aggarwal[1], Vibhuti Nautiyal[2], Garima Joshi[3], Nishit Galhotra[4]
[1,2,3,4] UG Students, SOC, DIT UNIVERSITY,
Dehradun, Uttarakhand, INDIA

Abstract: It is difficult for credit card firms to detect malicious activities like fraudulent transactions which cause its users to make payments from their accounts without their knowledge for the items that they did not purchase leading them to financial loss. As the world is moving towards digitalization the use of digital money has also increased which has also led to a rise in fraud associated with them parallelly. There are several methods applied to stop fraudulent activities but fraudsters keep on trying to find new ways and methods and always come up with unique ideas to break the security mechanism to commit fraudulent transactions making billions of losses to banks and credit card users globally. Therefore, there is a great demand for a technique for detecting credit card made fraudulent transaction that not only prevents it but also accurately and efficiently anticipates before it happens. This paper uses and explains various techniques for detecting credit card fraud, conducts a thorough analysis of both the existing models and the proposed model, and then conducts a comparison of these techniques based on achieved accuracy, false alarm rate, and detection rate.

Keywords:- *Random Forest, Logistic Regression, Decision Tree, SVM (Support Vector Mechanism), False Alarm Rate (FAR), Decision Rate.*

I. INTRODUCTION

As the world is moving towards digitalization, the use of digital money, and Internet Banking has become very common, any individual who is eligible can easily get a credit card issued from their bank to make any kind of online transactions, a credit card is a very thin plastic card that includes unique credit card number, cardholder’s name, signature, CVC code and validity information of the card, these information is required to make any online transaction, however with increase of credit card users, the credit card fraudulent activities has also increased parallelly. Today, banks, retail readers, ATMs, and online Internet banking systems all read information from credit cards. Its security relies on both the plastic card's physical security and the confidentiality of the credit card number, which is of the utmost significance. Credit card fraud essentially refers to any activity carried out with the intention of deceiving the card's owner as well as the bank which issued it in order to gain personal information for other fraud activities. To prevent such transaction, we need a powerful detection system that combat such activities in their initial stages before they become successful.

Fraud detection is a procedure that identifies and stops scammers from making money in dubious ways. It is a collection of actions performed to expose and thwart fraudsters' attempts to gain money or property fraudulently. The building of a model that will produce the best outcomes in detecting and avoiding the incidence of fraudulent transactions is referred to as detecting credit card theft with machine learning.

The work of detecting fraud is quite challenging; there are many characteristics that must be chosen and categorized, and the categorization of these parameters determines the effectiveness of any detection system. Furthermore, the current models can only determine the possibility of a transaction being fraudulent based on the analysis of user behaviours and activities. They attempt to identify patterns in the way users spend their money and evaluate if a transaction is legitimate or not.

Credit card fraud happens usually when there are:

1. **Clone Transactions: -** As the name suggests it refers to a duplication of a transaction. It is an easy way to copy all the information from any existing transaction and are frequently a well-liked technique for doing transactions that resemble the real thing.
2. **Account Theft: -** It often occurs when a person's private information, such as login credentials, the answer to a secret question, their birthdate, or any other information that is confidential, is taken by the culprit, who can then use it to carry out money transactions.
3. **False Application Fraud: -** As discussed above, account theft is generally coupled with application fraud. It signifies a fake account that is often referred to as one that has been applied for using another individual's name and identity.
4. **Credit Card Skimming (electronic or manual): -** Skimming a credit card refers to producing an unauthorized copy of a credit card using a skimmer, a device that reads and copies information from the original card. Using skimmers, fraudsters may copy or duplicate card numbers and other account information, preserve it, and then sell it to other criminals. Both manually and electronically are capable of being used.
5. **Account Takeover: -** It is one of the most commonly and widely used fraud technique, here fraudsters send deceptive calls and emails to cardholders, the messages they send feels genuine as if they were sent by the bank or any other official body, such messages are used to get and stealing a person's credentials, bank account numbers, and other confidential data, CVC code or

birthdate in order to perform financial profit by the fraudsters.

II. DIFFERENT TECHNIQUES FOR DETECTING CREDIT CARD FRAUD

Since almost all fraudulent transactions comply with a related pattern, we can classify transactions as fraudulent or legitimate using any pattern recognition algorithm, like *SVM (“Support Vector Machine”), LR (“Logistic Regression”), ANN (“Artificial Neural Network”), “Naive Bayesian Network”, KNN (“K-Nearest Neighbor”), “Random Forest”, “Hidden Markov Models”, a “Fuzzy-Logic-Based Systems”,* and *“Decision Trees”*. In our proposed model we have used the following techniques-

1. **Support Vector Machine: -** This is one of the well-known statistical learning approaches that has been proven to be highly successful in a range of classification tasks, even it may be utilized for regression issues as well. It is one of the supervised learning algorithms, where dataset is split into distinct classes using a hyperplane whose dimensions rely on the features, and the data points that are closest to the hyperplane within each class are those that belong to that class.
2. **Logistic Regression: -** When the dependent variable is categorical, it is another often- employed strategy. Clustering often uses logistic regression, and as a transaction is processed, it looks at the values of its characteristics to determine whether or not it should be committed. A supervised classification process known as logistic regression explains the connection between predictors that may be continuous, binary, or categorical.
3. **Decision Tree: -** It is among the effective computational tools used for making classification and prediction, it builds a tree like structure comprising of internal nodes, where each branch denotes the results of a test on an attribute, and a class label is held by each leaf node. It uses DFS or BFS techniques to recursively divide a dataset, and it stops when every element has been given a certain class.
4. **Random Forest: -** It is also one of the supervised learning algorithms that create and merge multiple decision trees into one forest, here the major goal is to work on the collection of decision models rather than relying only on a single model to improve accuracy. The major difference that lies between the “*Decision Tree”* and the “*Random Forest”* is that the decision tree produces a single model using the whole dataset, whereas the random forest builds several models using attributes from the dataset that are randomly chosen. This is the major justification for using the random forest model instead of the decision tree model.

III. LITERATURE REVIEW

R. M. Jamail Esmaily [1] [2015] In their paper have present an anomaly detection approach based on an ANN (*“Artificial Neural Network”*) and *“Decision Tree”*. The approach is divided into two stages. A multilayer neural network is used to categorize the data after a decision tree is used to create a brand new dataset. There are very few false detections in this two-level method.

Yashvi Jain, Shripriya Dubey, Namrata Tiwari, and Sarika Jain [2] [2019] thoroughly examined several machine learning techniques including ANN. They discovered that “*Artificial Neural Networks*” provide more exact results than *“Decision Tree”*, *“Logistic Regression”*, “*Support Vector Machine*” and *“K- Nearest Neighbor”* technologies.

M. Ramya, K. Anandh Raja and S. Ajith Kumar[11] [2020] observed multiple commonly used fraud detection methods and concluded that by using an API module and predictive analytics the user could be notified in real-time.

S P Maniraj, Shadab Ahmed, Aditya Saini, and Swarna Deep Sarkar [4] [2019] propose a unique technique for detecting fraudulent transactions by using various anomaly detection algorithms.

E. Duman and Y. Sahin [5] [2019] claimed that the decision tree strategy outperforms the SVM approach in answering the issue.

N. Malini and M. Pushpa [6] [2017] employed KNN and outlier detection to improve outcomes in fraud detection scenarios. The main objective was to reduce false alarms and raise the rate of fraud detection.

A.S.Malini, J.M Shajitha Banu, M.I Sharmila Fathima [7] [2022] They used Isolation Forests along with Area Under Precision-Recall Curve and observed an accuracy of 98.72%.

Kartik Madkaikar, Preity Parab, Manthan Nagvekar, Riya Raikar, and Supriya Patil [8] [2021] compared the implementations of multiple Classification techniques and observed Gradient Boosting to be the best with an accuracy of 95.90%.

Anuruddha Thennakoon, Shalitha Mihiranga, Chee Bhagyani, Sasitha Premadasa and Nuwan Kuruwitaarachchi [9] [2019] They addressed four main frauds in real-world transactions using a series of ML models where the highest accuracy observed was 91% using SVM.

Naresh Kumar Trivedi, Umesh Kumar Lilhore , Sarita Simaiya, and Sanjeev Kumar Sharma [10] [2020] tested multiple supervised learning algorithms and observed that Random Forest gave the maximum accuracy of 94.99%.

IV. METHODOLOGY

➢ *Dataset Description*

The dataset that we have used for our analysis is publicly available in the format of a CSV file, obtained from Kaggle. The dimensions of the dataset are 284807 rows and 31 columns. The dataset is a collection of 284807 real-world transactions, out of which 492 transactions are fraudulent. The dataset details are in the form of numerical variables produced by Principal component analysis transformation, with the columns ranging from V1 to V28. These variables provide information on the different features of a user's credit card transactions. The only unmodified features in the dataset are 'Time', 'Amount' and 'Class'. The 'Class' contains data in 0 and 1 format where 0 stands for valid credit card transactions and 1 stand for fraudulent credit card transactions.

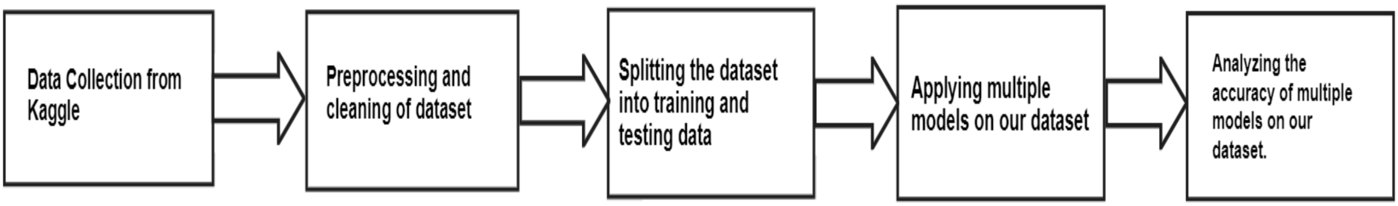

Data Collection from Kaggle
Preprocessing and cleaning of dataset
Splitting the dataset into training and testing data
Applying multiple models on our dataset
Analyzing the accuracy of multiple models on our dataset.

Fig 1 Flow Diagram

The dataset is obtained from Kaggle. After this we further analyse the dataset and perform the relevant pre-processing to make the dataset appropriate for our machine learning models and removed the unwanted feature time from our dataset. After cleaning of dataset, we split our dataset into testing and training data. We then test and train our dataset for different models including SVM, Logistic Regression, Decision Trees and Random Forest. Then we also obtain the confusion matrix after applying all the models. At last, we analyse and compare the accuracies of all these models.

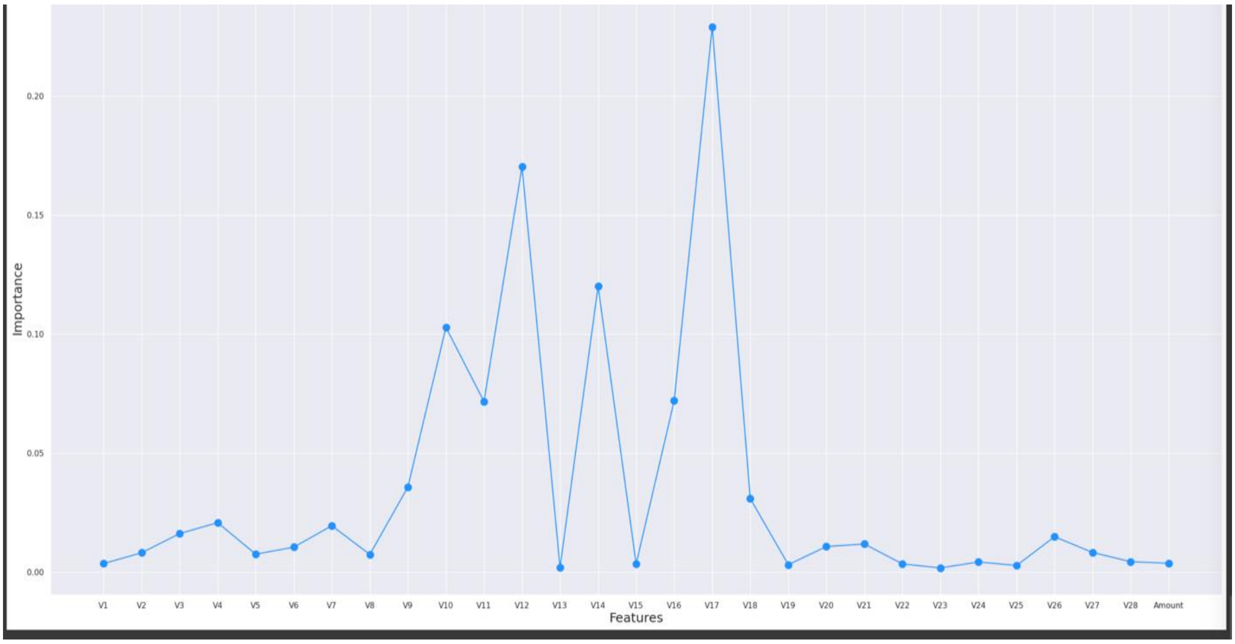


Fig 2 Relation between Features and Importance using Random Forest Model.

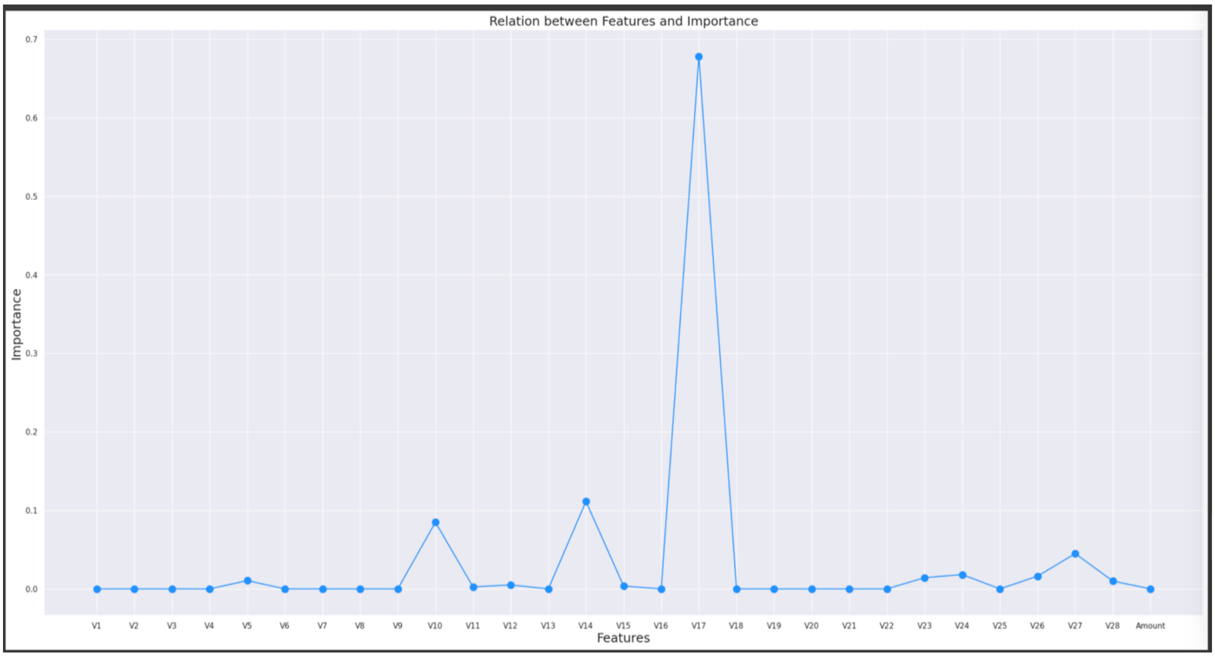


Fig 3 Relation between Features and Importance using Decision Tree

- *Comparative Analysis*

We compute the “true positive”, “false positive”, “true negative”, and “false negative” values produced by a system, also known as the confusion matrix, and use them as quantitative measures to assess the performance of various models in order to compare these methods.

The "true positive" (TP) rate is the proportion of transactions in the dataset that were both fraudulent and classified as such. The number of genuine transactions in the dataset that the system incorrectly identified as fraudulent transactions is known as a "false positive" (FP), while the number of fraudulent transactions in the dataset that the system incorrectly identified as legitimate transactions is known as a "false negative" (FN). The number of transactions in the dataset that were both lawful and accurately categorized as being genuine is known as the "true negative" (TN). Metrics employed in evaluation for our model are:

1. The percentage of transactions that the model accurately classifies is known as accuracy. It is among the most popular and widely applied assessment measures. Equation of accuracy (ACC) is given as: -

ACC = (TN + TP)/ (TP + FP + FN + TN)

2. Detection Rate also known as Precision is the proportion of transactions in the sample that the model correctly identified as either legitimate or unauthorized. Equation of Detection Rate (DR) is given as: -

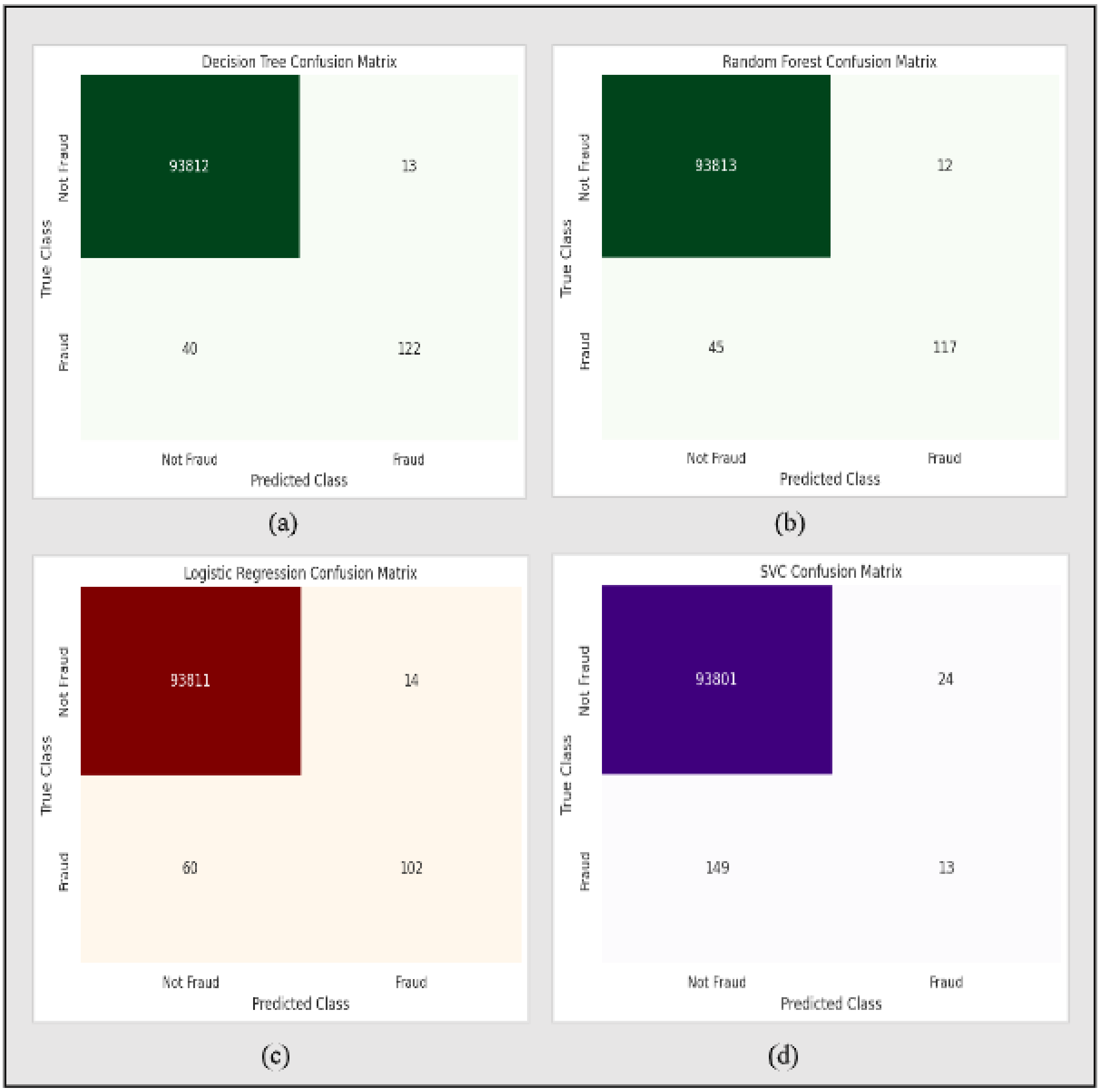
DR = TP / (TP + FP)

3. False Alarm Rate measures that how many were wrongly classified as fraudulent out of total instances classified as fraudulent. Equation of False Alarm (FAR) is given as: -

FAR = FP/ (FP+TN)

The performance of every machine learning model on a set of test data is outlined in a matrix called a *“confusion matrix”*. It is frequently used to assess the effectiveness of categorical label prediction algorithms, which try to predict a category label for each input occurrence.

The below fig 4 shows the confusion matrix obtained by the system after applying Logistic Regression(a), Random Forest(b) similarly it also shows the confusion matrix for Decision Tree(c) and SVM(d) model representing variety values of “True Positives” (TP), “False Positives” (FP), “True Negatives” (TN), and “False Negatives” (FN) generated by the above-mentioned techniques on the test data.

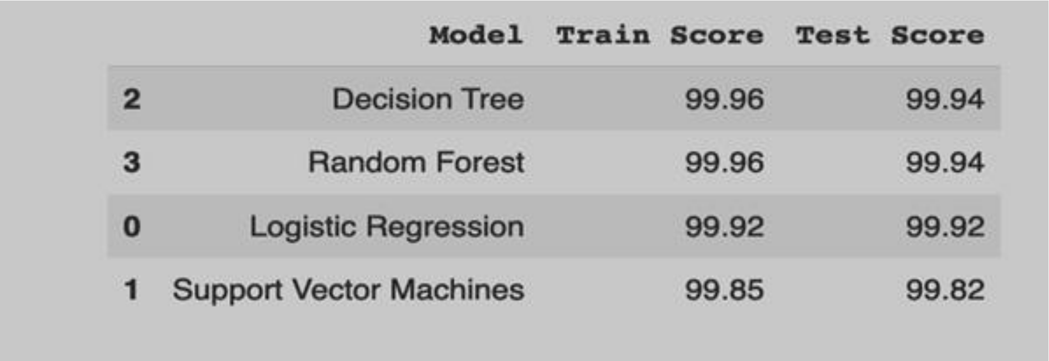


(a) (b)

(c) (d)

Fig. 4 Confusion Matrix

As discussed above after obtaining the confusion matrix, we perform the comparison between model accuracies, the below figure fig 5 shows the training and testing scores for the different methods that we have implemented in our model.



	Model	Train Score	Test Score
2	Decision Tree	99.96	99.94
3	Random Forest	99.96	99.94
0	Logistic Regression	99.92	99.92
1	Support Vector Machines	99.85	99.82

Fig. 5 Model Accuracy

Table 1 Comparative Analysis

Techniques	Accuracy	Detection Rate (Precision)	False Alarm Rate (FDR)
Support Vector Machine (SVM)	99.81	99.97	0.158
Random Forest	99.93	99.96	0.047
Decision Tree	99.94	99.99	0.042
Logistic Regression	99.92	99.98	0.063

The above Table 1 shows the comparison between the accuracy, detection rate also known as precision and false alarm rate values that we have achieved using different techniques, it can be clearly identified that we have achieved highest accuracy in decision tree (99.94 %) compared to SVM (99.81), random forest (99.93), and logistic regression (99.92). Similarly, we have also achieved high detection rate (99.99 %) and lowest false alarm rate with decision tree (0.042) as compared to other models.

V. CONCLUSION AND FUTURE SCOPE

In this project, we have implemented different techniques in order to get the best accuracy for the fraud detection. As we can see from the Table 1.1, where we have compared different techniques on different grounds, Decision. Tree has the best accuracy amongst the implemented techniques. It also has low false alarm rate and high precision Rate.

Although there are different techniques to find the card fraud detection but none are able to detect completely. The models generally detect after fraud has been committed, each techniques works best for a particular environment, for e.g., Decision tree works best for the already processed and sampled data, whereas logistic regression gives best result on raw and unsampled data. Hence the best solution for this problem is to use hybrid techniques in order to nullify the environment constraints and get the better performance.

ACKNOWLEDGEMENT

We would like to extend our deepest appreciation to Mr. Ashok Kumar (Assistant Professor, SOC, DIT UNIVERSITY) who has been our guide and mentor and who made this project possible. The advice and guidance he provided enabled us to get through each phase of conducting this paper. We would also like to add our severe thanks to DIT University, the institution that gave us both the opportunity and the resources we needed to finish this research project.

REFERENCES

[1]. R. M. Jamail Esmaily, “Intrusion detection system based on Multilayer perceptron neural networks and decision tree,” in International Conference on Information and Knowledge Technology, 2015.

[2]. Jain, Y. & Tiwari, N. & Dubey, S. & Jain, Sarika. (2019). “A comparative analysis of various credit card fraud detection techniques” in International Journal of Recent Technology and Engineering. 7. 402-407.

[3]. Siddhartha Bhattacharyya, Sanjeev Jha, Kurian Tharakunnel, J. Christopher Westland, “Data Mining for credit card fraud: A comparative study,” Elsevier, vol. 50, no. 3, pp. 602613, 2011.

[4]. S P Maniraj, Aditya Saini, Shadab Ahmed, Swarna Deep Sarkar, 2019, Credit Card Fraud Detection using Machine Learning and Data Science, International Journal of Engineering Research & Technology (IJERT) Volume 08, Issue 09 (September 2019).

[5]. Y. Sahin and E. Duman, “Detecting Credit Card Fraud by Decision Trees and Support Vector Machines,” Int. Multi-conference Eng. Computer Science, vol. I, pp. 442–447, 2011.

[6]. N. Malini and M. Pushpa, "Analysis on credit card fraud identification techniques based on KNN and outlier detection," 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), Chennai, 2017, pp. 255-258.

[7]. A.S.Malini, J.M Shajitha Banu, M.I Sharmila Fathima, “Credit Card Fraud Detection Using Machine Learning”, June 2022, IJIRT, Volume 9 Issue 1, ISSN: 2349-6002.

[8]. Kartik Madkaikar, Manthan Nagvekar, Preity Parab, Riya Raikar, Supriya Patil, “Credit Card Fraud Detection System”, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-10 Issue-2, July 2021.

[9]. Anuruddha Thennakoon, Chee Bhagyani, Sasitha Premadasa, Shalitha Mihiranga, Nuwan Kuruwitaarachchi, “Real-time Credit Card Fraud Detection Using Machine Learning”, 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)

[10]. Naresh Kumar Trivedi, Sarita Simaiya, Umesh Kumar Lilhore, Sanjeev Kumar Sharma, “An Efficient Credit Card Fraud Detection Model Based on Machine Learning Methods”, International Journal of Science and Technology Vol. 29, No. 5, (2020), pp. 3414 – 3424.