# Credit Card Fraud Detection Using Machine Learning as Data Mining Technique

DR ARUNA GAWADE[1], DR. NILESH RATHOD[2], HRIDAY SHETTY[3]
[1, 2, 3] *AI & ML, DJSCE, Mumbai University, Mumbai, India*

***Abstract*— Credit card fraud poses a significant threat to financial institutions, businesses, and consumers, necessitating advanced detection mechanisms to safeguard against unauthorized transactions. This research explores innovative approaches to credit card fraud detection using machine learning and data analytics. The study leverages a comprehensive dataset sourced from real-world transactions to develop and evaluate the effectiveness of fraud detection models. The research focuses on feature selection, model optimization, and the integration of emerging technologies in enhancing the accuracy and efficiency of detection systems. Results indicate promising performance, with the proposed models exhibiting notable success in identifying fraudulent activities. The findings contribute to the ongoing discourse on bolstering the resilience of financial systems against the evolving landscape of credit card fraud. The implications of the research extend to industry stakeholders, offering insights into refining existing detection strategies and inspiring future developments in the pursuit of heightened security measures.***

## I. INTRODUCTION

In recent years, the proliferation of electronic transactions and the widespread use of credit cards have brought unprecedented convenience to consumers and businesses. However, this digital transformation has also given rise to a significant and persistent challenge—credit card fraud. Fraudulent activities not only jeopardize the financial stability of individuals but also impose substantial economic burdens on financial institutions and merchants. As technology evolves, so do the methods employed by fraudsters, necessitating continuous advancements in detection mechanisms.

In the rapidly evolving landscape of the digital era, the prevalence of credit card fraud has surged to unprecedented levels, presenting a formidable challenge to the security of financial transactions. As electronic payment systems become integral to everyday life, the sophistication of fraudulent activities has reached new heights, necessitating advanced and adaptive fraud detection mechanisms.

### *1.1 Background: Rising Prevalence of Credit Card Fraud*

The pervasive digitization of financial transactions has given rise to a parallel escalation in credit card fraud. Cybercriminals exploit vulnerabilities in the interconnected web of electronic payment systems, perpetrating fraudulent activities that range from the conventional theft of card information to sophisticated cybercrimes such as identity theft and data breaches. The increasing reliance on digital payment methods and the expansion of online commerce have created an expansive canvas for fraudsters to exploit, emphasizing the urgency of fortifying financial systems against these ever-evolving threats.

### *Highlighting the Importance of Effective Fraud Detection Mechanisms*

As the prevalence of credit card fraud intensifies, the importance of effective fraud detection mechanisms cannot be overstated. These mechanisms serve as the first line of defense, safeguarding individuals, businesses, and financial institutions from the financial and operational repercussions of unauthorized transactions. Timely and accurate detection not only mitigates direct financial losses but also preserves the trust and confidence essential for the continued growth of electronic payment systems.

### *1.2 Significance of the Study: Economic and Social Consequences*

The economic and social consequences of credit card fraud are profound, affecting individuals, businesses, and society at large. For individuals, falling victim to fraud entails immediate financial losses, coupled with the potential for long-term damage to credit histories and personal trust in financial systems. Businesses, especially those engaged in e-commerce, face operational disruptions, financial burdens, and reputational damage. The broader societal impact includes eroding trust in electronic payment systems,

hindering the widespread adoption of digital financial technologies.

*Illustrating the Significance of Advancing Detection Systems*

Considering the escalating economic and social consequences of credit card fraud, advancing detection systems becomes imperative. Enhancing the resilience of these systems is not only a technological necessity but also a crucial step in maintaining trust in electronic payment platforms. The significance of this study lies in its potential to contribute innovative insights that can elevate the effectiveness and efficiency of credit card fraud detection, thereby fortifying the foundations of electronic financial transactions.

*1.3 Research Problem: Challenges and Limitations*

Against this backdrop, this research seeks to address the inherent challenges and limitations present in current credit card fraud detection methods. Traditional approaches, while effective to a certain extent, often struggle to keep pace with the dynamic and adaptive nature of contemporary fraud techniques. Identifying these challenges is fundamental to formulating adaptive solutions that can proactively counter evolving fraud tactics.

*Acknowledging the Need for Adaptive Solutions*

The need for adaptive solutions is paramount, recognizing that fraudsters continuously refine their methods to exploit emerging vulnerabilities. Static and rule-based systems are proving inadequate in the face of this evolving landscape. Therefore, this research acknowledges the pressing requirement for innovative, adaptive solutions that can stay one step ahead of fraudsters, ensuring the continued integrity of electronic payment systems.

*1.4 Research Objectives: Developing Advanced Models*

This research is driven by a set of interrelated objectives designed to address the identified challenges:

1. Develop advanced credit card fraud detection models that leverage cutting-edge technologies, such as machine learning and data analytics.
2. Systematically evaluate the effectiveness and efficiency of the proposed models in real-world scenarios.
3. Contribute insights that go beyond the immediate scope of detection, aiming to enhance the overall resilience of financial systems against the persistent threat of fraud.

*1.5 Research Question or Hypothesis: Guiding the Investigation*

At the core of this research lies a pivotal inquiry that guides the investigation: How can advanced machine learning and data analytics techniques be harnessed to develop more robust credit card fraud detection models? This question sets the stage for a comprehensive exploration, seeking not only to address the intricacies of contemporary fraud but also to contribute transformative insights to the broader field of cybersecurity and financial technology.

## II. LITERATURE REVIEW

The landscape of credit card fraud detection has witnessed significant advancements in recent years, fueled by the escalating sophistication of fraudulent activities. This literature review synthesizes existing research, focusing on credit card fraud detection methodologies, identifies gaps in current approaches, and explores the pivotal role of machine learning (ML) and data analytics in countering these challenges.

*1. Traditional Approaches to Credit Card Fraud Detection:*

Credit card fraud is a growing problem in the financial industry, with the potential to cause significant financial losses to both customers and financial institutions. As a result, there has been a significant amount of research in recent years on developing effective fraud detection systems. These systems rely on a combination of statistical techniques, machine learning algorithms, and deep learning models to identify fraudulent transactions. One of the most commonly used approaches for credit card fraud detection is rule-based systems. These systems use predefined rules to identify transactions that are deemed suspicious. (Shah, Akshat & Makwana, Yogeshvari. (2023). Credit Card Fraud Detection.)

*2. Challenges and Limitations in Current Approaches:*

The method explained above is not perfect in the true sense. Here are some of the challenges that complicate the fraud detection process -

1. Changing fraud patterns over time — This one is the toughest to address since the fraudsters are always in the lookout to find new and innovative ways to get around the systems to commit the act. Thus, it becomes all-

important for the deep learning models to be updated with the evolved patterns to detect. This results in a decrease in the model's performance and efficiency. Thus, the machine learning models need to keep updating or fail their objectives.

2. Class Imbalance — Practically only a small percentage of customers have fraudulent intentions. Consequently, there's an imbalance in the classification of fraud detection models (that usually classify transactions as either fraudulent or non-fraudulent) which makes it harder to build them. The fallout of this challenge is a poor user experience for genuine customers, since catching the fraudsters usually involves declining some legitimate transactions.

3. Model Interpretations — This limitation is associated with the concept of explainability since models typically give a score indicating whether a transaction is likely to be fraudulent or not — without explaining why.

4. Feature generation can be time-consuming — Subject matter experts can require long periods of time to generate a comprehensive feature set which slows down the fraud detection process. (Medium.com, 2019)

*3. The Rise of Machine Learning in Fraud Detection:*

Machine learning plays a vital role for detecting the credit card fraud in the transactions. For predicting these transactions banks make use of various machine learning methodologies, past data has been collected and new features are been used for enhancing the predictive power. The performance of fraud detecting in credit card transactions is greatly affected by the sampling approach on data-set, selection of variables and detection techniques used.

Dataset of credit card transactions is collected from kaggle and it contains a total of 2,84,808 credit card transactions of a European bank data set. It considers fraud transactions as the "positive class" and genuine ones as the "negative class." The data set is highly imbalanced, it has about 0.172% of fraud transactions and the rest are genuine transactions. The author has been done oversampling to balance the data set, which resulted in 60% of fraud transactions and 40% genuine ones. The three techniques are applied for the dataset and work is implemented in R language. The performance of the techniques is evaluated for different variables based on sensitivity, specificity, accuracy, and error rate. The result shows of accuracy for logistic regression, Decision tree and random forest classifier are 90.0, 94.3, 95.5 respectively. The comparative results show that the Random Forest performs better than the logistic regression and decision tree techniques. (Lakshmi S V S S1 ,Selvani Deepthi Kavila2, 2018)

*4. Unsupervised Learning and Anomaly Detection:*

The usage of credit card has increased dramatically due to a rapid development of credit cards. Consequently, credit card fraud and the loss to the credit card owners and credit cards companies have been increased dramatically. Credit card Supervised learning has been widely used to detect anomaly in credit card transaction records based on the assumption that the pattern of a fraud would depend on the past transaction. However, unsupervised learning does not ignore the fact that the fraudsters could change their approaches based on customers' behaviors and patterns. In this study, three unsupervised methods were presented including autoencoder, one-class support vector machine, and robust Mahalanobis outlier detection. The dataset used in this study is based on real-life data of credit card transaction. Due to the availability of the response, fraud labels, after training the models the performance of each model was evaluated. The performance of these three methods is discussed extensively in the manuscript. For one-class SVM and auto encoder, the normal transaction labels were used for training. However, the advantages of robust Mahalanobis method over these methods is that it does not need any label for its training. (Rezapour Mashhadi, Mohammad Mahdi. (2019). Anomaly Detection using Unsupervised Methods: Credit Card Fraud Case Study. International Journal of Advanced Computer Science and Applications. 10. 10.14569/IJACSA.2019.0101101.)

*5. Deep Learning and Neural Networks:*

Fraud detection systems support advanced detection techniques based on complex rules, statistical modelling, and machine learning. However, alerts triggered by these systems still require expert judgement to either confirm a fraud case or discard a false positive. Reducing the number of false positives that fraud analysts investigate, by automating their detection with computer-assisted techniques, can lead to significant cost efficiencies. Alert reduction has been achieved with different techniques in related fields like intrusion detection. Furthermore, deep learning has been used to accomplish this task in other fields. In our paper, a set of deep neural networks have been tested to measure their ability to detect false positives, by processing alerts triggered by a fraud detection system. The performance achieved by each neural network

setting is presented and discussed. The optimal setting allowed to capture 91.79% of total fraud cases with 35.16% less alerts. Obtained alert reduction rate would entail a significant reduction in cost of human labor, because alerts classified as false positives by the neural network would not require human inspection. (@article{San_Miguel_Carrasco_2020, doi = {10.1109/access.2020.3026222}, url = {https://doi.org/10.1109%2Faccess.2020.3026222}, year = 2020, publisher = {Institute of Electrical and Electronics Engineers ({IEEE})}, volume = {8}, pages = {186421--186432}, author = {Rafael San Miguel Carrasco and Miguel-Angel Sicilia-Urban}, title = {Evaluation of Deep Neural Networks for Reduction of Credit Card Fraud Alerts}, journal = {{IEEE} Access} })

*6. Integration of Data Analytics:*

Fraud comes in various forms and continues to evolve as technology advances. Some of the common types of fraud include credit card fraud, insurance fraud, and identity theft. In credit card fraud, malicious actors use stolen credit card information to make unauthorized transactions. Insurance fraud involves individuals exaggerating or fabricating insurance claims to receive undue compensation. Identity theft occurs when someone unlawfully obtains and uses another person's personal information for financial gain or to commit other fraudulent activities.

Explanation of How Data Analytics Can Be Applied to Each Type of Fraud:

- Credit Card Fraud: Data analytics can play a crucial role in detecting and preventing credit card fraud. By analyzing transaction data, patterns, and user behavior, machine learning algorithms can identify anomalies and potentially fraudulent activities. For instance, if a credit card is suddenly used for transactions in different geographical locations within a short time span, the system can trigger an alert. Data analytics can also establish baselines for typical spending habits, allowing deviations to be easily spotted. Additionally, real-time monitoring can help block suspicious transactions before they are completed.
- Insurance Fraud: Data analytics can aid in tackling insurance fraud by analyzing historical claims data and identifying irregularities. Advanced algorithms can flag claims that deviate from typical patterns, such as unusually frequent claims from a specific policyholder. Text mining techniques can be employed to scan claim descriptions for keywords associated with fraudulent claims. By integrating external data sources, such as medical records or accident reports, analytics systems can cross-reference information and identify inconsistencies.
- Identity Theft: Data analytics can contribute to identifying identity theft through anomaly detection and behavioral analysis. By analyzing login patterns, geographic locations, and device usage, systems can identify unusual activities that may indicate unauthorized access. Machine learning models can be trained to recognize behaviors that differ from an individual's historical patterns. Moreover, data analytics can be employed to correlate multiple data sources to detect instances where stolen identities are being used for financial transactions or fraudulent applications. (iabac.org, 2023)

*7. Gaps and Future Directions:*

By extracting out the article' objectives and conclusions, we can recognise trends, conduct gap analysis, determine future research. As a result, to identify the gaps and define the next direction of future research should take, based on the article's objectives and conclusions, we conducted a summary analysis.

We identified research gaps by investigating unexplored or infrequently studied algorithms. In addition, we found supervised learning as the most prevalent learning technique and SMOTE as the most prevalent oversampling technique. Many researchers focused on supervised techniques such as LR, RF, SVM, and NN

We examine the trend of the reviewed article. In addition, we compare the models created over the three years to determine and evaluate which techniques recently garnered more attention. This also assist, to identify the gaps so that future research will be able to address them in their own work. First, we examined the distribution of the chosen article by the publication year. In year 2019 (47 articles), 2020 (70 articles), and 2021 (64 articles). Significant difference existed between the years 2019 and 2020, the number of published articles for credit card fraud detection increased

(23 articles). However, there was no notable difference between 2020 and 2021 (six articles)

(Eyad Abdel Latif Marazqah Btoush1 , Xujuan Zhou1 , Raj Gururajan1,2, Ka Ching Chan1 , Rohan Genrich1 and Prema Sankaran3, 2023)

## III. METHODOLOGY

### 1. Data Collection

The dataset is sourced from ULB Machine Learning Group and description is found in. The dataset contains credit card transactions made by European cardholders in September 2013. This dataset presents transactions that occurred in two days, consisting of 284, 807 transactions. The positive class (fraud cases) make up 0.172% of the transactions data. The dataset is highly unbalanced and skewed towards the positive class. It contains only numerical (continuous) input variables which are because of a Principal Component Analysis (PCA) feature selection transformation resulting to 28 principal components. Thus, a total of 30 input features are utilized in this study. The details and background information of the features cannot be presented due to confidentiality issues. The time feature contains the seconds elapsed between each transaction and the first transaction in the dataset. The 'amount' feature is the transaction amount. Feature 'class' is the target class for the binary classification and it takes value 1 for positive
case (fraud) and 0 for negative case (non fraud).

### 2. Data Preprocessing

Generally, data transformation and data reduction are referred to as data preprocessing phase, where the raw data is cleaned and Journal of Telecommunication, Electronic and Computer Engineering 26 e-ISSN: 2289-8131 Vol. 10 No. 1-4 transformed into appropriate forms (or standardization) to be evaluated and fed into machine learners. Data transformation process involves activities such as normalization, smoothing, aggregation, attributes construction and generalization of the data. While data reduction is to reduce the number of attributes such as data cube aggregation, removing irrelevant attributes and principal component analysis. For instance, during data transformation, the format of transaction date and time were standardized into a uniform state so that it was identical to machine learners to interpret it as date and time attributes. Then, Principal Component Analysis technique was employed to detect the anomaly transactions. Principal Component Analysis is a method to transform the correlated variables into a smaller number of uncorrelated attributes called Principal Components. The objective of applying the method was to identify and reduce the dimensionality of the dataset and discover new meaningful underlying attributes. The advantage of Principal Component Analysis is during reducing the dimensions of the data using eigenvector, the losses to the information of the data are insignificant. Furthermore, the losses could be trace back by decompressing the eigenvalue.

### 3. Feature Selection

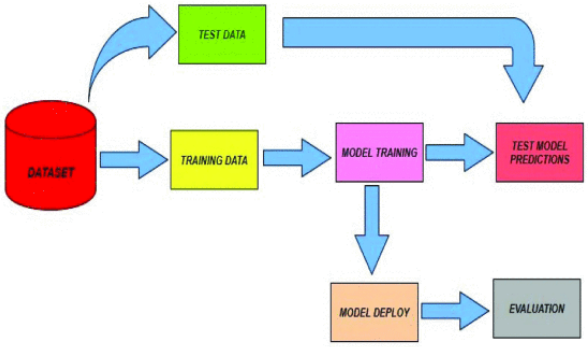
The basis of credit card fraud detection lies in the analysis of cardholder's spending behaviour. This spending profile is analysed using optimal selection of variables that capture the unique behaviour of a credit card. The profile of both a legitimate and fraudulent transaction tends to be constantly changing. Thus, optimal selection of variables that greatly differentiates both profiles is needed to achieve efficient classification of credit card transaction. The variables that form the card usage profile and techniques used affect the performance of credit card fraud detection systems. These variables are derived from a combination of transaction and past transaction history of a credit card. These variables fall under five main variable types, namely all transactions statistics, regional statistics, merchant type statistics, time-based amount statistics and time-based number of transactions statistics.

The variables that fall under all transactions statistics type depict the general card usage profile of the card. The variables under regional statistics type show the spending habits of the card with considered the geographical regions. The variables under merchant statistics type show the usage of the card in different merchant categories. The variables of time-based statistics types identify the usage profile of the cards with respect to usage amounts versus time ranges or frequencies of usage versus time ranges. Most literature focused on cardholder profile rather than card profile. It is evident that a person can operate two or more credit cards for different purposes. Therefore, one can exhibit different spending profile on such

cards. In this study, focus is beamed on card rather than cardholder because one credit card can only exhibit a unique spending profile while a cardholder can exhibit multiple behaviours on different cards.

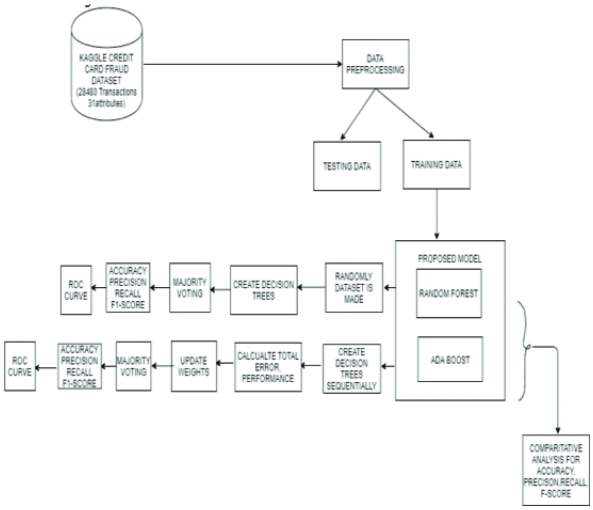
4. Model Selection

The main aim of this paper is to classify the transactions that have both the fraud and non-fraud transactions in the dataset using algorithms like that the Random Forest and the Adaboost algorithms. Then these two algorithms are compared to choose the algorithm that best detects the credit card fraud transactions. The process flow for the credit fraud detection problem [Figure.3.]includes the splitting of the data, model training, model deployment, and the evaluation criteria.




Process Flow

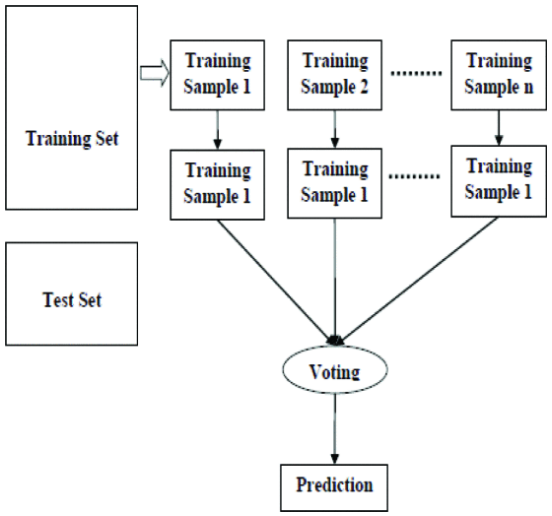
The detailed architecture diagram for the credit card fraud detection system [Figure. 4.] includes many steps from gathering dataset to deploying model and performing analysis based on results. In this model we take the Kaggle credit card fraud dataset and pre-processing is to be done for the dataset. Now to prepare the model we have to split the data into the training data and the testing data. We use the training data to prepare the Random Forest and the Adaboost models. Then we develop both the models. Finally, the accuracy, precision, recall, and F1-score is calculated for bot the models. Finally the comparison of the credit card fraud transactions more accurately.




Architecture Diagram

A. Random Forest Algorithm

The Random Forest algorithm [Figure. 5] is one of the widely used supervised learning algorithms. This can be used for both regression and classification purposes. But, this algorithm is mainly used for classification problems. Generally, a forest is made up of trees and similarly, the Random Forest algorithm creates the decision trees on the sample data and gets the prediction from each of the sample data. Then Random Forest algorithm is an ensemble method. This algorithm is better than the single decision trees because it reduces the over-fitting by averaging the result.




Random Forest Algorithm

*Steps for Random Forest Algorithm*

1. Take the Kaggle credit card fraud dataset that is trained and randomly select some of the sample data.
2. Using the randomly created sample data now creates the Decision Trees that are used to classify the cases into the fraud and non-fraud cases.
3. The Decision Trees are formed by splitting the nodes, the nodes which have the highest Information gain make it as the root node and classify the fraud and non-fraud cases.
4. Now the majority vote is performed and the decision Trees may result in 0 as output which includes that these are the non-fraud cases.
5. Finally, we find the accuracy, precision, recall, and F1 -score for both the fraud and non-fraud cases.

*Random Forest algorithm*

Algorithm Random Forest:

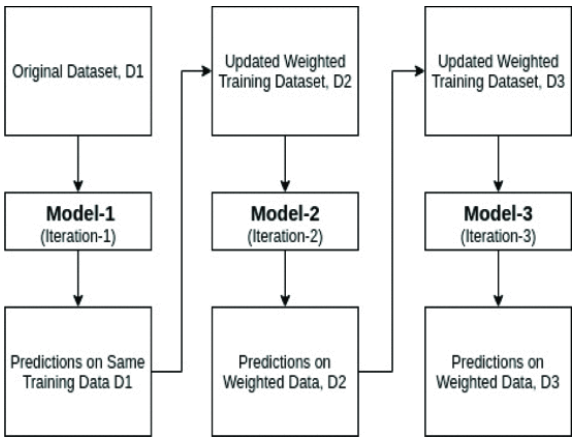
```
Algorithm Random Forest :
To generate c classifiers:
        For i=1 to c do
        Randomly select the training data D with
        replacement to produce Di
        Create a root node N containing Di and cell
Build Tree(N)
End for
Majority Vote

Build Tree(N)
        Randomly select x% of all the possible splitting
features in N
        Select the features F that has the highest Information
        A gain for further splitting
        Gain (T,X)=Entropy (T)-Entropy(T,X)
        Now to calculate the entropy we use,
        E(S) = ∑_{i=1}^{c}(−Pi log Pi)
        Create f child nodes
        For i=1 to f do
                Set contents f N to Di
                Call Build Tree(Ni)
        End for

End
```

B. Adaboost Algorithm

Boosting is one of the ensemble techniques. This algorithm is used to build strong classifiers from weaker classifiers. This can be done by building a strong model by using a weak model in the series. Initially, a model is built from the training data. Then the second model is built from the first model by correcting the errors that represent in the model that is created before. This is a repetitive process and is continued until either the maximum number of models is added or the complete training dataset is predicted correctly. Adboost was one of the most successful boosting algorithms that were developed for the binary classification.




Adaboost Algorithm

The short name for Adaboost is adaptive boosting. It is best used with weak learners. This Adaboost boosting technique [Figure. 6] combines the multiple weak classifiers into a strong classifier. Adaboost algorithm can be used with short decision trees. The way the Adaboost is created is such that initially at first the nodes are created and the tree is made, then the performance of the tree on each of the instances is checked. Also, a weight is assigned. The training data that is hard to predict is the one that gives more weight. The Adaboost algorithm is a powerful classifier that works well on both the basic and complex problems. The disadvantage of this algorithm is that this algorithm is mostly sensitive to noisy data. This algorithm is also sensitive to outliers.

Steps for Adaboost Algorithm

1. The Kaggle credit card fraud dataset is taken and is trained. Randomly select some of the sample data.
2. Using the randomly created sample data now creates the decision trees sequentially for classifying the fraud and non-fraud cases.
3. The decision trees are formed initially. This can be done by splitting the node based on which has the highest information gain, make it as the root node, and classify the fraud and non-fraud cases.
4. Now calculate the error rate, performance, and update the weights of the fraud and non-fraud transactions that are incorrectly classified.

5. Now majority vote is performed and the decision trees may result as output which indicates the nonfraud cases.
6. The decision trees may output 1 which indicates that it is a fraud case.
7. Finally, we find the accuracy, precision, recall, and F1-score for both the fraud and non-fraud cases.

Adaboost Algorithm

Algorithm Adaboost:

Algorithm Adaboost :
IINPUTdataset
Initialize weights, wl(n)=l/n
Create a decision tree
Select the one that has the lowest Entropy
If Incorrectly classified
Calculate Total Error (TE)= sum of up incorrectly Classified sample weights
Calculate Performance, $P = \log\frac{1-TE}{TE}$
For each
Incorrectly classified, increase weights:
Weights incorrect =old weight $*e^{p}$
Correctly classified, decrease the weights:
Weight correct =old weight $*e^{-p}$
Normalized weight of each sample:

$$\text{Normalized weight} = \frac{updated\ weight}{sum\ of\ updated\ wight}$$

End for
End if

5. Evaluation Metrics

To compare various algorithms, we need to evaluate metrics like accuracy, precision, recall, and F1-score. The confusion matrix is also plotted. The confusion matrix is a 2*2 matrix. The matrix contains four outputs which are TPR, TNR, FPR, FNR. Measures such as sensitivity, specificity, accuracy, and error-rate can be derived from the confusion matrix Then we that best suit to detect the credit card fraud.

The output of the confusion matrix is

True Positive Rate, which can be defined as the number of fraudulent transactions that are even classified by the system as fraudulent.

True Negative Rate, which can be defined as the number of legitimate transactions that are even classified as legitimate by the system.

False Positive Rate, which can be defined as a number of the legal transactions which are wrongly classified as fraud.

False Negative Rate is defined as the transactions that are fraud but are wrongly classified as legal.

The Receiver Operating Characteristics curve is created by plotting the TPR against the FPR. This can be done at various thresholds. ROC curve is a graph in which the FPR is the horizontal axis and the TPR is the vertical axis. The graph under the ROC curve is the AUC.

## IV. DATA ANALYSIS

The confusion matrix and the ROC curve is plotted for both the algorithms. The dataset, when applied for different algorithms, gives different outputs. Firstly we apply the dataset for the random forest model and the results are as below:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 1.00 | 1.00 | 1.00 | 93825 |
| 1 | 0.95 | 0.77 | 0.85 | 162 |
| accuracy | | | 1.00 | 93987 |
| macro avg | 0.97 | 0.89 | 0.93 | 93987 |
| weighted avg | 1.00 | 1.00 | 1.00 | 93987 |

Output for Random Forest

The evaluation criteria are explained fer for that of the fraud cases.

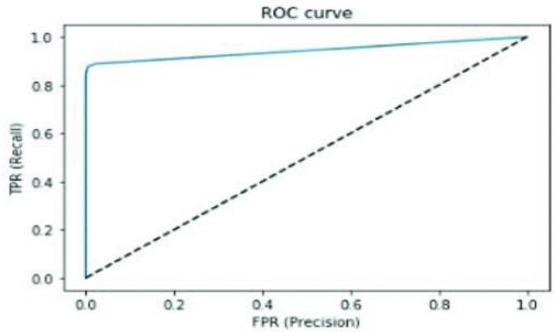
```
Confusion Matrix on train data
[[190490      0]
 [     0    330]]

Confusion Matrix on test data
[[93818    37]
 [    7   125]]
```

Confusion Matrix for Random Forest

The confusion matrix [Figure.8] shows us that for the train data the true positives are 190490 and false positives are 0, the true negatives are 0 and the false

negatives are 330. For the test data, the true positives are 93818 and false positives are 37, the true negatives are 7 and the false negatives are 125.




```
Area under curve (AUC):  0.9429434888303349
```

ROC curve for Random Forest

Now the dataset is applied for the Adaboost algorithm The results are obtained like that of the Random Forest Algorithm

```
Accuracy = 0.9990743400683073
              precision    recall  f1-score   support

           0  0.99938202 0.99969091 0.99953644     93825
           1  0.78195489 0.64197531 0.70508475       162
```

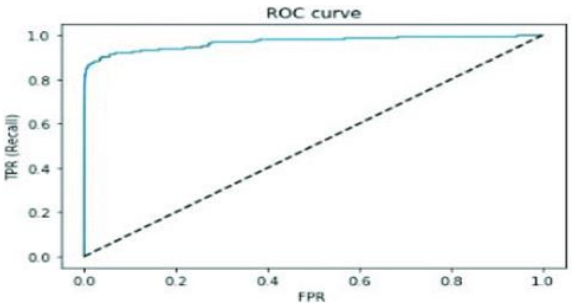
Output for Adaboost

The evaluation criteria shows us that the evaluation criteria like the precision, recall, and F1-score differ less in the case of the non-fraud cases and differ greatly in those of the fraud cases.

```
Confusion Matrix on train data
[[190464    120]
 [    26    210]]
Confusion Matrix on test data
[[93811    65]
 [   14    97]]
```

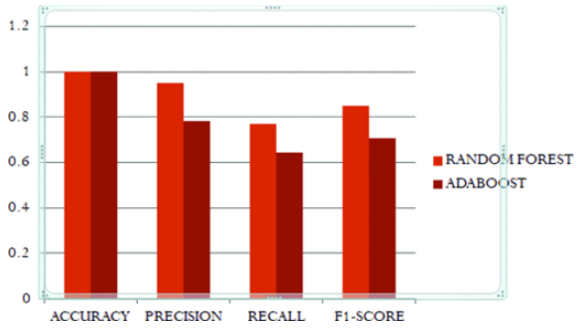
Confusion Matrix for Adaboost

The confusion matrix [Figure.11] shows us that for the train data the true positives are 190464 and false positives are 120, the true negatives are 26 and false negatives are 201. For the test data, the true positives are 93811 and false positives are 65, the true negatives are 14 and false negatives are 97.




ROC curve for Adaboost

Now the comparison of the random forest and the Adaboost algorithms is shown [Figure.12]. The two algorithms have the same accuracy but the precision, recall, and the F1-score of the two algorithms differ. The random forest algorithms have the highest precision, recall, and F1-score.




Comparision of Algorithms

## CONCLUSION

Even though there are many fraud detection techniques we cannot say that this particular algorithm detects the fraud completely. From our analysis, we can conclude that the accuracy is the same for both the Random Forest and the Adaboost algorithms. When we consider the precision, recall, and the F1-score the Random Forest algorithm has the highest value than the Adaboost algorithm. Hence, we conclude that the Random Forest Algorithm works best than the Adaboost algorithm to detect credit card fraud.

## REFERENCES

[1] R. Sailusha, V. Gnaneswar, R. Ramesh and G. R. Rao, "Credit Card Fraud Detection Using Machine Learning," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India,

2020, pp. 1264-1270, doi: 10.1109/ICICCS48265.2020.9121114. keywords: {Machine learning algorithms;Forestry;Credit cards;Control systems;Electronic commerce;Random forests;credit card fraud;fraudulent activities;Random Forest;Adaboost;ROC curve},

[2] https://journalofbigdata.springeropen.com/articles/10.1186/s40537-022-00573-8#citeas

[3] https://dl.acm.org/doi/10.1145/3377049.3377113