Random Forest Classifier for Credit Card Fraud Detection

Vijay Gaikwad[1], Suvarna Mane[2], Disha Chandak[3], Om Pardeshi[4], Tanaya Barawkar[5], Sagar Yadav[6], Gopal Dhanpalwar[7]
[1,2,3,4,5,6,7]*Vishwakarma institute of technology, Pune, India*

Abstract: **Credit card fraud (CCF) is a straightforward and appealing target. Online sites as well as E- commerce have enlarged their payment options, as risk of online fraud is increasing rapidly. Researchers began using various machine learning (ML) algorithms to detect and analyses online transaction fraud as fraud rates increased. This paper presents a random forest-based model to detect fraudulent transactions by analyzing customers' historical transaction details and extracting behavioral patterns. Cardholders are divided into groups based on the volume of their transactions. Then, using a sliding window method, the transactions performed by cardholders are aggregated from various categories. The behavioral patterns of the various groupings are then derived. The random forest (RF) classifier shows the greatest accuracy and hence proved to be one of the most excellent ways for detection/prediction of frauds. As a result, a feedback mechanism is implemented to address the issue of notion drift. The proposed model provides high accuracy of 99.99% with precision of 93% and recall of 73%. The proposed model provides better performance than isolation forest algorithm as well logistic regression and support vector machine.**

Keywords: **fraud detection, random forest algorithm, machine learning, credit card fraud detection, isolation forest algorithm.**

I. INTRODUCTION

CCF is becoming more prevalent each day. Online and physical transactions both can be used to commit CCF. But however physical cards are very important in offline transactions, or else virtual cards can be used for online transactions for unlawful or fraudulent activity. As a result of these credit card fraud operations, many fraudulent transactions may occur if there is lack of information and knowledge [2]. In order to conduct any type of transactions, scammer need careful details like the number of credit card (CC), funds holder with number, whatever necessary users' qualification. Fraudsters must run off the user's CC to perform offline purchases, but in order to execute online transactions, the user's identity and online data are required which should be run off by the scammers. As a result, CCF has emerged as a serious concern in today's technology society, with a significant impact on bank transactions. But it's very difficult for the users and the banking officials to predict different fraud transactions, hence resulting in the loss of confidential data [2]. There are many replicas for recognizing fraud transactions which are basically based on transaction behavior, and these methods perhaps classified into two types: supervised learning and unsupervised learning algorithms. They employed methods including Support Vector Machine, Cluster Analysis, logistic regression and Nave Bayer's Classification in the existing system to decide the accuracy of crooked activities. The purpose of this work is to utilize the Random Forest Algorithm (RFA) to decide the correctness of fraudulent transactions. Fraud is a lawbreaker crime committed by an unofficial person who cheats the innocent. CCF occurs when thieves obtain the cardholder's personal information and use it in an unlawful way, such as through SMS or phone calls. This credit card theft may also be carried out through the use of fraudulent operating system. CCF is detected in the following way: the buyer provides the appropriate qualification in sequence to conduct any credit card transaction, and the transaction should only be accepted after it has been thoroughly investigated for possible fraud [3]. To do so, the transaction data is sent to the verification module, which determines if the transaction is fraudulent or not. Any transaction classified as crooked is refused. Otherwise, the deal is accepted.

A. Problem Statement

Now a days most of the people uses CC to purchase products that they require, since this avoids use of cash. CC are being utilized more frequently to meet demands, and the scams linked with them are increasing causing financial loss to the people. In order to avoid this loss, a model is required which can detect fraudulent transactions with greater precision and accuracy.

B. Objective

- The major goal is to find out a crooked transaction in CC financial transactions.
- When unsupervised learning and supervised learning were compared, the supervised learning algorithm proved to be the most effective strategy concentrating on accuracy.

II. RELATED WORK

There have been numerous ways proposed in previous studies to find best ways to discover the scam, ranging from supervised method to unsupervised methods to cross-breed approaches; this necessitates learning the technologies involved in CCFD as well as a thorough comprehension of the various types of CCF. Scam trends develop over time, creating latest types of fraud, making it a hot topic among researchers. The remainder of this section follows through individual ML algorithms/models, and different fraud identifying systems that have been employed in fraud identifying. The matter that arose throughout the research has been investigated to be able to develop an effective ML replica in the future [12]. The RFA is a supervised ML technique that break down credit card transactions using a decision tree and thereafter leverages a prediction model to calculate performance. The proposed technique has a 90 percent accuracy rate [13]. eila provides an approach for gathering profiles that takes advantage of the intrinsic design in installment of transactions, with fraud detection taking place online [15]. They investigate and analyse several algorithms in order to detect CFF, like the SVM and RF, and conclude that RF performs the best in the aggregation process. The aggregated approach, however, failed to identify a scam in actual in this investigation. Navanushu Khare and Saad Yunus Sait presented their decision-tree research, random forests, logistic regression and SVMs in 2018 [16]. They used a highly forked dataset to create this type of dataset. Accuracy, sensitivity, specificity, and precision are used to assess performance. The accuracy of LR is 97.7%, Decision Trees is 95.5 percent, RF is 98.6 percent, and SVM classifier is 97.5 percent, as per the results. They concluded that, among the other algorithms, hence it's clear that in case of accuracy RF is the greatest algorithm for detecting scam. They also came to the conclusion that the SVM method has a file discrepancy issue and it doesn't perform well in detecting CCF. ML technique like SVM, Nave Bayes, k-Nearest Neighbor, etc. are engaged in current regime, and others have used the isolation forest approach. Few people have utilized the random forest method to detect credit card fraud.

III. SYSTEM ARCHITECTURE

In order to detect and classify CCF transactions, a system based on random forest algorithm is proposed as shown in fig. 1. The credit card dataset contains all of the information on credit cards. However, the amount and transaction time were evaluated while analyzing and pre-processing the dataset. The data cleaning technique is the next step, which involves evaluating the dataset and removing all duplicate and null values. Data partitioning divides the CC databank into 2 parts: a trained dataset and a testing dataset. Following that, a confusion matrix is generated using the RFA. The confusion matrix is used to do the performance analysis. This performance study will yield a CCFD accuracy of more than 90%.

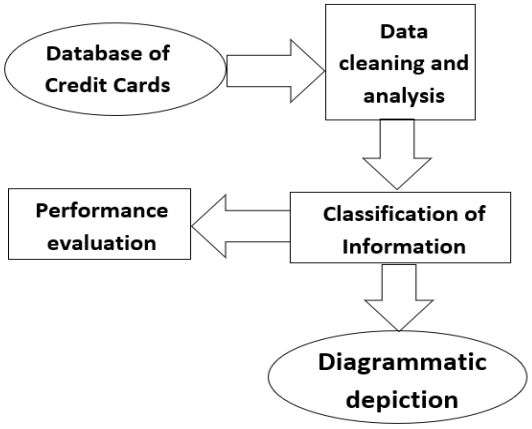

Fig. 1. System Architecture of CCFD

A. Data pre-processing

The dataset is preprocessed using data cleaning approach. The first step is loading of the dataset. Data cleansing and normalization is performed. Data splitting divides the dataset into two portions, one for amount and one for transaction time. The model is trained and evaluated on these two portions. Finally, the system determines whether or not the transaction is valid.

B. Programming language used

Python is used as the best programming language in ML to develop the proposed system. Python has been the new trend in recent years due to its ease of use, ability

to interpret, as having object-oriented approach. It has a lot of packages and libraries for ML.

C. Random Forest algorithm

RF is also known as Random Decision Forest (RDF), and it is applied for categorization, retrogression, and different functions that require many decision trees to be constructed. This RDF is built on supervised learning and has the benefit of being able to be utilized for both classification and regression. In comparison to all other current regime, the RFA provides superior accuracy, as it creates many decision trees and blends them together. Even the missing values will be handled by the random forest classifier, which will retain the correctness of a major percentage of the data. According to studies, random forest is widely used in various areas such as banking, medical, stock market, and health sector because it has excellent accuracy, when compared to any other algorithm. As a result, the idea of using random forest for credit card fraud came into existence and was implemented in the research, resulting in greater accuracy. Fig. 2 describes mechanism for majority voting (shown in blue circles) for the different features.

Regression problems:

When solving regression problems with the Random Forest Algorithm, make use of the mean squared error (MSE) to see how each elements data stream out

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2 \quad \ldots\ldots\ldots (1)$$

where N is the overall number of parameters,

f_i is the model's output value and

y_i is the proper data element quantity i.

This algorithm estimates the space between each node and the expected real benefit, allowing you to choose which branch is best for your woodland. The quantity of the periodic item for checking at a particular location is y_i .while f_i is decision trees returned value.

The output of RFA is estimated using equation (2)

$$RFfi_i = \frac{\sum_j norm\ fi_{ij}}{\sum_{j \in all\ features, k \in all\ trees} norm\ fi_{jk}} \quad \ldots\ldots\ldots (2)$$

Again, random forest makes use of Mean Absolute Error (MAE) while solving regression problems.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \mu| \quad \ldots\ldots (3)$$

where y_i = is the label for an instance,

N = is the quantity of instances,

μ = is the mean given by $\frac{1}{N}\sum_{i=1}^{N} y_i$

Classification problems:

When accomplishing RF build on categorization data, the Gini index is used as follows

$$Gini = 1 - \sum_{i=1}^{C}(pi)^2 \quad \ldots\ldots (4)$$

These techniques, which uses the class and expectation to compute Which branch has the greatest chance of appearing, is used to calculate the Gini per each datatype stem. pi shows the relative frequency, and C is the amount of classes available.

Entropy can also be used to connect the elements on decision tree so for this it uses a certain expectation.

$$Entropy = \sum_{i=1}^{C} - p_i * log_2\ (p_i) \quad \ldots\ldots (5)$$

D. Detection of CCF using RFA

Fig.3 shows detection of CCF using RFA, so When it comes to detecting CCF, the RFA boosts accuracy. The entire dataset will be collected and examined first. All duplicate values, as well as invalid, will be deleted from the dataset during the analysis process. The dataset will now be initialized to decide the accuracy of the output databank depending on the amount and transaction time. The databank will now be separated into 2 groups once it has been initialized into amount and transaction time. There are two types of data in the dataset: training data and test data. Hence a programmed named 'Kaggle' to classify datasets was utilized. 'Kaggle' is a free ML library written in Python that includes features such as categorization, retrogression, centering techniques, and other algorithms that work with Python. now the RFA is used to process the dataset once it has been initialized. The initialized dataset will be evaluated again using the RFA, and a CM will be generated. In the CM, the data will be separated into 4 hinder: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The databanks are constantly divided till whole data has been confirmed. All of the divided data will now be examined, and the results will be displayed as distinct graphs. These different graphs will simply provide little information about the outcome. As a result, in order to improve accuracy, RFA is used to extract all charts of the dataset and provide required values with greater accuracy when contrast with other techniques.

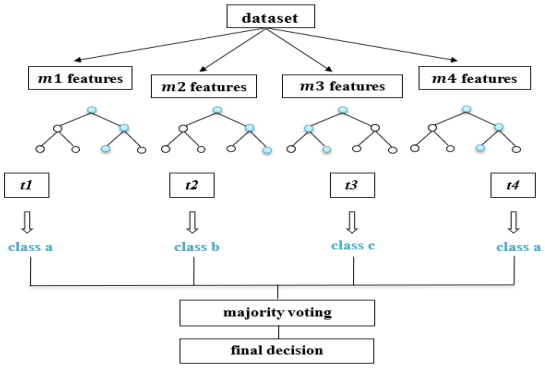

Fig. 2. Random forest algorithm

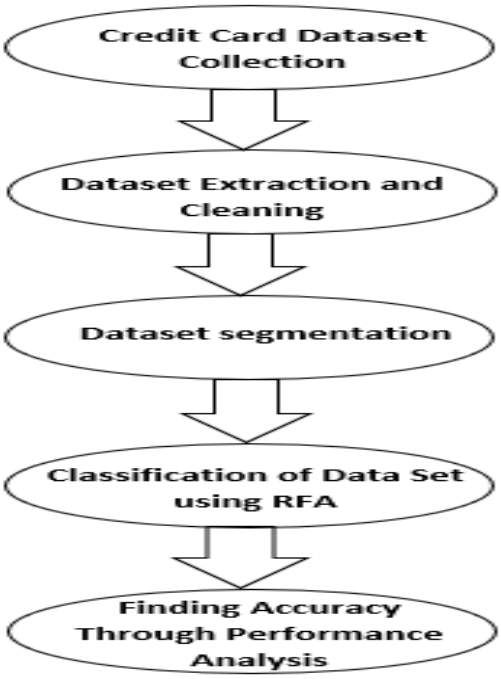

Fig. 3. Detection of CCF through RFA

IV RESULTS

A. Dataset

Both the dataset and the source code were obtained from Kaggle. The files contain CC transactions done by European cardholders in September 2013 which contains around 56809 transactions splitted into 31 columns and 5 rows. The original features have been replaced with V1, V2, due to some confidentiality concerns. V28 columns are the outcome of applying PCA transformation on the original ones. 'Time' and 'Amount' are the only factors that have not been changed by PCA. The answer variable, 'Class,' takes the value 1 in the case of scam and 0 otherwise.

Time:

The time between this transaction in seconds and the dataset's initial transaction.

Amount:

Amount of transaction

Class:

1 if the transaction is fraudulent, 0 otherwise

B. Histograms for displaying dataset characteristics

To see if there were any anomalous parameters, histograms for all of the parameters were plotted. The end result is displayed below.

C. Correlation matrix

The correlation matrix graphically depicts how features link to one another and can aid in predicting which features are most important for the prediction. From the Heatmap it's clear that the majority of the features do not correlate with one another, but some do have a positive or negative association with one another. The attributes "V2" and "V5", for example, are significantly adversely related to the characteristic "Amount." There is also a connection between "V20" and "Amount." This aids in the comprehension of the data. The accuracy, precision, recall, and F1-score of the isolated forest algorithm vs random forest method were calculated, and the results are shown below.

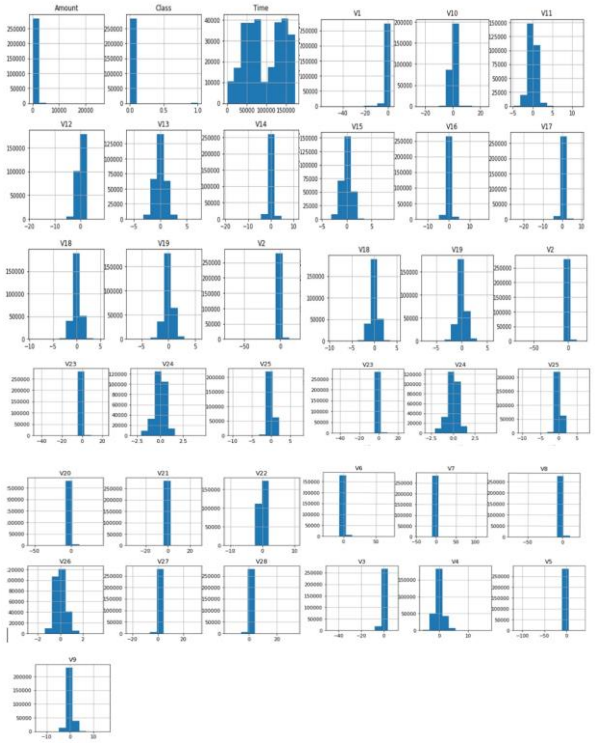

Fig. 4. Histograms of all 31 parameters

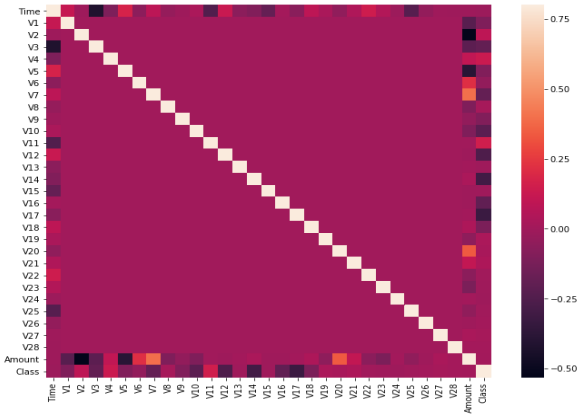

Fig. 5. Correlation matrix

D. Confusion matrix

A CM is a diagram that displays how well a Categorization system is applied to a collection of test data that has known true values. Precision, recall, and accuracy are calculated once the last result is estimated using the confusion matrix. It has two types of classes: actual and projected. These characteristics influence the confusion metrics:

True Positive (TP): When both numbers are positive, the result is 1.

True Negative (TN): When both numbers are negative, the result is 0.

False Positive (FP): When the true class is 0 but the non-true class is 1.

False Negative (FN): This occurs when the true class is 1 and the false class is 0.

This is how precision is defined:

Precision = TP / Final outcome

$$Precision = TP/ (TP+ FP) \ldots\ldots\ldots (6)$$

This is how recall is defined:

Recall = TP / expected outcome

$$Recall = TP/ (TP + FN) \ldots\ldots\ldots (7)$$

This is how accuracy is defined:

$$Accuracy = (TP+ TN)/ total \ldots\ldots\ldots (8)$$

This is how F1-Score is defined:

$$F1\text{-}Score = 2* (Precision * Recall) / (Precision + Recall) \ldots\ldots\ldots (9)$$

Hence below the fig. 6 and 7 is the confusion matrix which displays the results of the isolated forest method and the RFA, and based on the confusion matrix different parameters have been calculated. As a result, it's clear that the random forest is the best algorithm for detecting CCF. As confusion matrix provides number of true detections as well as false detections, this performance parameter is very much useful in evaluating the model outcome. The confusion matrix not only provides information about true positives but also provides information about the ability of the model to correctly reject the non-desired results. The fig. 8 shows how different algorithms compare in terms of accuracy. These graphics are created by comparing different study papers. After examining a variety of methodologies, it has been determined that RF offers the highest accuracy, estimated to be around 99.99 percent. As a result, it turned out to be the best CCFD algorithm. The fig. 9 shows how different algorithms compare in terms of precision. These graphics are created by comparing different study papers. After examining a variety of methodologies, it has been determined that RF offers the highest precision, of around 93% percent. As a result, it turned out to be the best CCFD algorithm. The fig. 10 shows how different algorithms compare in terms of recall value. These graphics are created by comparing different study papers. After examining a variety of methodologies, it has been determined that even the recall value of RF is greater than any other algorithm, which is around 73%. The fig. 11 shows how different algorithms compare in terms of F1- Score. These graphics are created by comparing different study papers. After examining a variety of methodologies, it has been determined that RF offers the highest F1-Score value that is 0.84 and hence proved to be one of the best algorithms of CCFD.

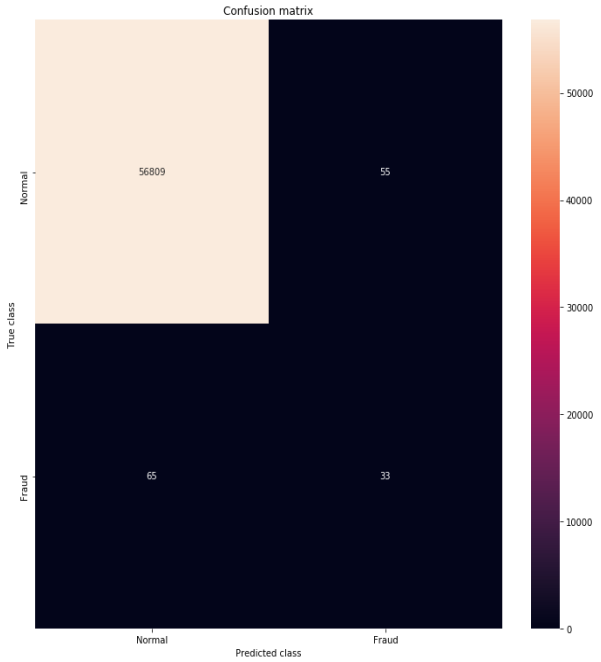

Fig. 6. Confusion matrix of isolation forest algorithm

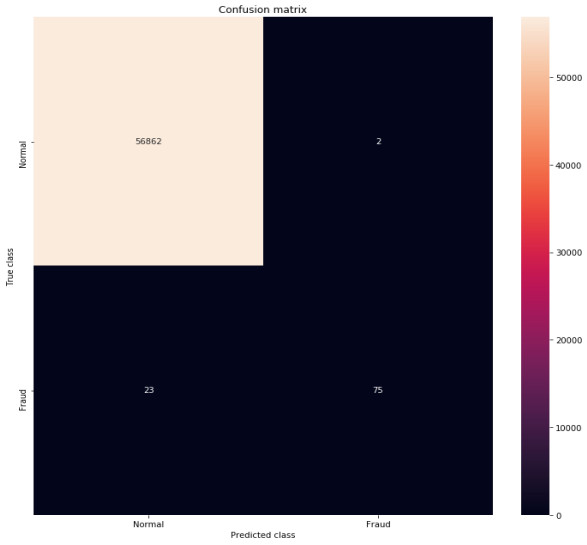

Fig. 7. Confusion matrix of random forest algorithm

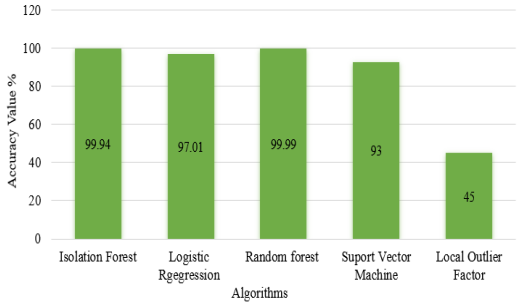

Fig. 8. Comparison between various algorithms based on accuracy

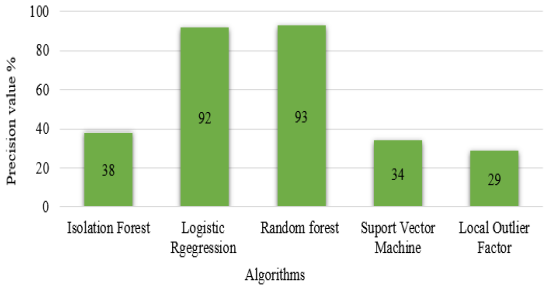

Fig. 9. Comparison between various algorithms based on precession

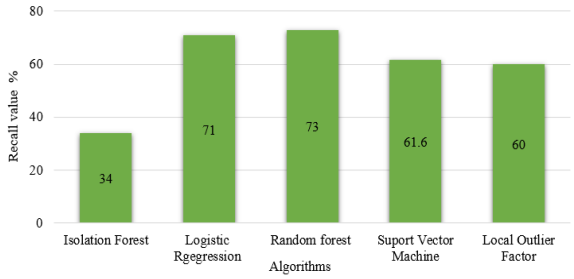

Fig. 10. Comparison between various algorithms based on recall

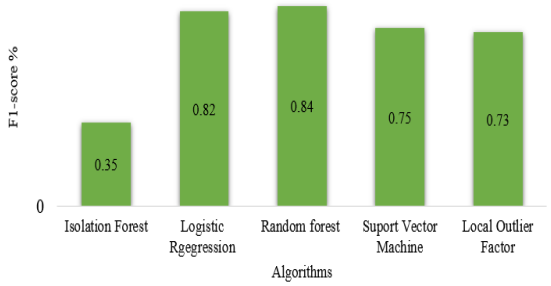

Fig. 11. Comparison between various algorithms based on F1-score

Table I Comparison between isolation forest algorithm and random forest algorithm

Algorithm	Accuracy	Recall	F1-Score	Precision
Isolation Forest algorithm	99.78 %	34%	35%	38%
Random Forest algorithm	99.99%	73%	0.84%	93%

V CONCLUSION

In spite of the fact that there are various fraud detection techniques, the proposed model detects fraudulent transactions with 99.99 % accuracy and precision with 93%. Based on the findings, it can be inferred that the Random Forest approach is more accurate than other ML algorithms such as isolation Forest, Logistic regression, support vector machine and local outlier factor in detection of fraudulent transactions related to the credit card. With multiple performance parameters such as accuracy, precision, recall, and F1-score, the Random Forest algorithm outperforms over other algorithms. It is thus observed that the RFA proves its model efficiency for identification of credit card theft which is very much useful in e-commerce business.

VI FUTURE SCOPE

In order to further improve the performance of the proposed system in terms of recall and F1-score, an approach based on deep learning architecture will be useful for CCFD.

REFERENCE

[1] Vaishnavi Nath Dornadulaa, Geetha Sa (2019). "Credit Card Fraud Detection using Machine

Learning Algorithms". https://doi.org/10.1016/ j.procs.2020.01.057.

[2] M. Suresh Kumar, V. Soundarya, S. Kavitha, E.S. Keerthika, E. Aswini (2019). "Credit Card Fraud Detection Using Random Forest Algorithm". https://doi.org/10.1109/iccct2.2019.8824930.

[3] Asha RB, Suresh Kumar KR (2021). "Credit card fraud detection using artificial neural network" science direct". https://doi.org/10.1016/j.gltp.2021.01.006.

[4] Shiyang Xuan, Guanjun Liu, Zhenchuan Li, Lutao Zheng, Shuo Wang, Changjun Jiang (2018)."Random Forest for Credit Card Fraud Detection". https://doi.org/10.1109/icnsc.2018.8361343.

[5] Ruttala Sailusha, V. Gnaneswar, R. Ramesh, G. Ramakoteswara Rao (2020). "Credit Card Fraud Detection Using Machin". https://doi.org/10.1109/iciccs48265.2020.9121114.

[6] Lakshmi S V S, Selvani Deepthi Kavila (2018) "Machine Learning for Credit Card Fraud Detection System". https://doi.org/10.1109/iciccs48265.2020.9121114.

[7] Maja Puh, Ljiljana Brkic (2019). "Detecting Credit Card Fraud Using Selected Machine Learning Algorithms" https://doi.org/10.23919/mipro.2019.8757212.

[8] M. Zarcapoor and P. Shamsolmoali, (2015). "Application of credit card fraud detection: Based on bagging ensemble classifier". https://doi.org/10.1016/j.procs.2015.04.201.

[9] A.Gupta, D. Kumar, and A. Barve, (2017). "Hidden Markov Model based Credit Card Fraud Detection System with Time Stamp and IP Address". https://doi.org /10.5120/ijca2017914060.

[10] S. MAs, K. Tuyls, B. Vanschoenwinkel, and B. Manderick, "Machine Learning Techniques for Fraud Detection," Proc. 1st Int. naiso Congr. neuro fuzzy Technol., no. January, pp. 261 270, 2002.

[11] M. Zareapoor and P. Shamsolmoali,(2015)."Application of credit card fraud detection: Based on bagging ensemble classifier," https://doi.org/10.1016/j.procs.2015.04.201

[12] I. Sohony, R. Pratap, U. Nambiar, (2018), Ensemble learning for credit card fraud detection, in: Proceedings of the ACM India Joint International Conference on Data Science and Management of Data Association for Computing Machinery, https://doi.org/10.1145/3152494.3156815.

[13] M.S. Kumar, V. Soundarya, S. Kavitha, E.S. Keerthika, E. Aswini, (2019). "Credit card fraud detection using random forest algorithm, in: Proceedings of the 3rd International Conference on Computing and Communications Technologies (ICCCT)" https://doi.org/10.1109/iccct2.2019.8824930.

[14] Anuruddha Thennakoon, Chee Bhagyni, Sasitha Premadasa, Shalitha Mihiranga, Nuwan Kuruwitaarachchi (2019) "Real-time Credit Card Fraud Detection Using ML" https://doi.org/10.1109/confluence.2019.8776942

[15] Seyedhossein, Leila, and M. R. Hashemi. (2018) "Mining information from credit card time series for timelier fraud detection." https://doi.org/10.1109/icnsc.2018.8361343.

[16] Navanshu Khare, Saad Yunus, (2021). Credit Card Fraud Detect Ion Using Machine Learning Models and Collating Machine Learning Models, International Journal of Pure and Applied Mathemat ics Volume 118 No. 20 2018, 825-838 ISSN: 1314-3395, https://doi.org/10.30534/ijeter/ 2021/02972021.