Fraud Detection of Credit Card using Machine Learning Algorithm

Asiya Mariyam A[1], Agnus S[2], Dr. P Kavitha[3]

[1]Panimalar Engineering College, India

[2]Panimalar Engineering College, India

[3]Panimalar Engineering College, India

Abstract

Credit card fraud refers to the loss or theft of the data or the credit card. This frequently occurs when the card is used improperly or when an attacker uses the data for his own benefit. Supervised Learning algorithms are the applied algorithms. For the purpose of detecting fraudulent transactions, the detection ofcredit card fraud software has been set up. While detecting entirely of the bogus transactions is difficult, the objective hereis to cut down on inaccurate fraud categories. The employed algorithms are the algorithm of Random Forest, logistic regression algorithm and K-Nearest Neighbor algorithm. The three algorithms' results are examined using the metrics precision, F1-score, accuracy and recall. The ROC curve is plotted employing the confusion matrix as a foundation. The best algorithm for spotting fraud is one that is the most exact, precise, recallable, and F1-score.

Keywords: Credit Card Fraud, Random Forest Algorithm, Logistic Regression, K-Nearest Neighbor algorithm, ROC- Receiver Operating Characteristic Curve

1. Introduction

A fraudulent counterfeit is defined as an intentional deception that is committed for abenefit, most often a financial one. This is an unfair action that occurs more frequently every day. The use of credit and debit cards as a form of payment has sharply increased,which has in turn increased the number of credit card frauds. It's possible that the card Won't need to be physically shown while paying online. In these kinds of situations, card details may be the target of hackers or cyber criminals. Due to this kind of scam, hundreds of thousands of dollars goes missing in a single year. To bypass this issue, several algorithms have been devised and are being worked on. Different detection methods are being developed in order to solve this problem as efficiently as feasible.

Payments made with credit cards have become very widespread, yet they also have a distinctive set of challenges. The sharp increase in the use of credit and debit cards as a form of payment has led to a spike in fraudulent use of credit cards. The procedure of detecting forgery is fraught with difficulties. It takes only a very brief period, possibly between a few seconds and nanoseconds, to authorize or refuse a transaction. As a result, the process utilized to spot a scam must be incredibly quick and effective. The enormous number of comparable transactions that happen to be taking place at the same time presents another difficulty. As a result, it is difficult to track each transaction individually to spotfraud. Therefore, it is essential to develop a trustworthy identifying fraudulent activity system in order to make the distinction between a valid transaction and one that is fraudulent.

One can commit two types of fraudulent transactions with credit cards. Robbery of confidential data from the card itself, such as the number of the card, cvv code, kind of card, and others. The first is the physical larceny of the card. Before the cardholder is aware, a fraudster can use stolen credit card information to steal a sizable sum of money or make a sizable purchase. In order to identify between genuine and forged transactions, organizations use a number of machine learning approaches.

All authorized payments are examined by machine learning algorithms, which are thenused to recognize any that stand out as suspicious. Determining if the transaction was honest or dishonest, experts investigate these complaints and contact the cardholders. The computerized system gathers evidence from the researchers, which is employed to train and modify the algorithm in order to enhance its performance in spotting deception over time.

The objective of this paper is to assess which supervised machine learning model is most effective in identifying credit card fraud. This study evaluates an unbalanced dataset using a variety of models. Three supervised machine learning models are used. A dataset is evaluated using an algorithm of Random Forest, logistic regression algorithm and K-Nearest Neighbor algorithm using various predefined criteria. Accuracy, precision,
F1-score and ROC curve are the criteria used to analyze which algorithm works best for credit card detection.

The rest of the paper is split into different sections: Section 2 presents an overview of the existing concepts, Section 3 gives the description of how the system works and its methodologies, the study's results and a comparative analysis are discussed in Section 4, and the outcome is laid out in Section 5.

2. Related Works

1. A credit card fraudulent activity happens regularly and ends up costing a lot of money. Online credit card transactions now make up a sizable portion of all transactions conducted online, which has seen significant growth. As a result, banks and other financial organizations provide very valuable and in-demand credit card fraud detection programmes.

Fraudulent transactions can take many different forms and fall under several categories. The four primary fraud incidents in real-world transactions are the topic of this research. A variety of machine learning models are used to address each scam, and the optimal approach is ultimately chosen after examination. This review offers a thorough manual for choosing the best algorithm for the kind of frauds, and we provide an appropriate performance metric to explain the evaluation.
Real-time credit card fraud detection is another important essential topic that we cover in our project. To determine if a certain transaction is legitimate or fraudulent, we employ predictive analytics performed by the integrated machine learning models and an API module. We also evaluate a cutting-edge approach that successfully tackles the skewed distribution of data. According to a private disclosure agreement, the financial institution provided the data forour studies.

2. The primary issue in the current era is the identification of credit card fraud.
This is a result of the expansion of e-commerce platforms and online transactions.In most cases, credit card fraud occurs when the card is stolen andused for any unauthorized activity, or even when the fraudster utilizes the card's information for his own gain. In the modern world, there are several issues with credit cards. The credit card fraud detection technology was created to identify fraudulent actions. The primary emphasis of this research is machine learning algorithms. The Random Forest Algorithm and the

Adaboost Algorithm is employed. The two algorithms' outputs are based on
F1-score, accuracy, precision, recall, and other metrics. On the basis of the confusion matrix, the ROC curve is plotted. Comparing the Random Forest and Adaboost algorithms, the method with the highest accuracy, precision, recall, and F1-score is regarded as the ideal approach for use in fraud detection.

3. The term "credit card fraud" describes the actual loss of a credit card or the loss of private credit card data. For detection, a variety of machine learning techniques can be applied. This study presents many methods for categorizing transactions as fraudulent or legitimate. The research made use of a dataset for credit card fraud detection. SMOTE was employed for oversampling since the dataset was quite unbalanced. Additionally, a feature selection process was carried out, and the dataset was divided into training and test halves. Naive Bayes, Multilayer Perceptron, Random Forest, and Logistic Regression were the algorithms employed in the experiment. The outcomes demonstrate the excellent accuracy of each algorithm for detecting credit card fraud. The proposed model may be employed to find other abnormalities.

4. Financial crime is a significant issue that is only getting worse and has
Far-reaching effects. In order to identify credit card fraud in online transactions, data mining was crucial. Due to two main factors—first, the profiles of legitimate and fraudulent activity vary often, and second, credit card fraud datasets are extremely skewed—detection of Credit card fraud, a data mining challenge, becomes difficult. The dataset sampling strategy, variable choice, and detection method(s) employed all have a significant impact on the effectiveness of fraud detection in credit card transactions. The effectiveness of naive bayes, k-nearest neighbor, and logistic regression on highly skewed credit card fraud data is examined in this research. Financial fraud is a serious problem that is only getting worse and has far-reaching effects. In order to identify credit card fraud in online transactions, data mining was crucial.
Due to two main factors—first, the profiles of legitimate and fraudulent activity vary often, and second, credit card fraud data sets are extremely skewed—detection of credit card fraud, a data mining challenge, becomes difficult. The dataset sampling strategy, variable choice, and detection method(s)employed all have a significant impact on the effectiveness of fraud detection in credit card transactions. The effectiveness of naive bayes, k-nearest neighbor, and logistic regression on highly skewed credit card fraud data is examined in this research.

5. The quick involvement in transactional activities that are mostly based online generates false instances everywhere and causes significant losses to the personal and financial business. Even if there are many illegal actions taking place in commercial enterprises, fraudulent ecard activities are among the most common and upsetting to online customers. The patterns and features of suspicious and non-suspicious transactions were examined using data processing

Techniques supplemented by knowledge of normalization and anomalies. On the other hand, victimization classifiers employed machine learning (ML) approaches to detect suspicious and non-suspicious transactions automatically. The supervised based
Mainly categorization is discussed in this study. All classifiers outperformed results obtained before preprocessing the dataset by over 95.0% accuracy after utilizing normalization and Principal element Analysis to preprocess the dataset.

6. For customers to avoid spending for products they did not buy, credit card issuers must be able to recognize fraudulent credit card transactions. Data Science may be used to solve these issues, and coupled with machine learning, it is of utmost relevance. With the use of credit card fraud detection, this research aims to demonstrate the modeling of a data set using machine learning. Modeling previous credit card transactions using information from those that turned out to be fraudulent is part of the Credit Card Fraud Detection Problem. The validity of a new transaction is then determined using this approach. The goal here is to minimize inaccurate fraud categories while detecting 100% of the fraudulent transactions. A classic example of categorization is the detection of credit card fraud. The analysis and
Pre-processing of data sets, as well as the use of several anomaly detection techniques, such as the Local Outlier Factor and Isolation Forest algorithm, to PCA-transformed Credit Card Transaction data have been the main points of this approach.

7. In the current economic climate, using a credit card has become very prevalent. These cards make it possible for the userto make significant payments without having to carry a lot of cash. They have revolutionized cashless transactions and made it simple for customers to make payments of any kind. Although incredibly helpful, this electronic payment method has a unique set of dangers. The number of credit card scams is rising at a comparable rate to the number of users. An individual's credit card details may have been fraudulently obtained and may have been used in fraudulent purchases. To solve this issue, several machine learning algorithms may be used to gather data. This study compares many well-known supervised learning methodsfor identifying legitimate from fraudulent transactions.

8. Nowadays, digitization is becoming more and more popular due to how fluid,simple, and convenient ecommerce is. It spread quickly and became a simple method of payment. People prefer to payand purchase online due to time and transportation ease, among other factors.
Due to the extensive usage of
E-commerce, credit card theft has also significantly increased. Fraudsters attempt to abuse the card and the security of online transactions.
Therefore, it is crucial to stop scammersin their tracks. The primary goal is to protect credit card transactions so that users may use e-banking with confidence and simplicity. Fraud

involving credit cards can be spotted using an array of approaches, including Deep Learning, Naive Bayesian, SupportVector Machine (SVM), Artificial Immune System, K Nearest Neighbor, Data Mining, Logistic Regression, Decision Tree, Neural Network, GeneticAlgorithm, etc.

9. Credit cards have been the most popularform of payment in recent years. The development of technology and the increase in fraud cases necessitate the development of an algorithm for detecting fraud that can accurately identify and stop fraudulent conduct. This study effort recommends a variety of machine learning-based methods for classification, including the algorithm of Random Forest, logistic regression algorithm, and K-Nearest Neighbor algorithm to handle the severely disproportionate dataset. Last but not least, the study's precision, f1 score, accuracy, recallable, and Roc-auc score will be calculated.

10. Credit card usage has expanded as a result of the swift growth of electronic commerce systems. As credit cards become the main mode of payment, there is a boost in credit card fraud for all the purchases made online and over time. Economic fraud has dramatically increased due to the development of contemporary innovations and Superhighways for telecommunication on a global scale that cost billions of dollars annually. Genuine transactions mix in with the counterfeit ones, it might be challenging to consistently discover forgeries using standard

Pattern-matching tools. Therefore, all

Credit card providing institutions are obligated to put in place efficient systems for identifying fraud in an effortto bring down their losses. The use of machine learning, fuzzy logic, artificial intelligence, and data mining are used in many recent techniques.

3. Proposed System

The dataset utilized in the studies is described in this section. The three classifiers that are being investigated, namely Logistic Regression, k-Nearest Neighbor, and Random Forest algorithm. The generation of classifiers involves several steps: gathering data, cleaning it up, analyzing it, programming the classification algorithm, testing, etc. During the preprocessing stage, the data are fitted and sampled into a format that is useful. Two distinct collections of data distributions have been generated using the two the positive and negative scenarios in unison. The classification algorithms are generated and provided using the cleaned data during this phase. The performance comparison of the classifiers is looked at based on f1-score, recall, precision, accuracy, and ROC curve.

A.Dataset

The data provided consists of cardholders' payments made with credit cards from September 2013 across Europe. 274,897 transactions that occurred during the period of two working days are included in this dataset. The positive class (fraud cases) makes about 0.185% of the total value of transaction data. The dataset is severely prejudiced and out of balance in favor of thepositive class. Only constant variables for input from a feature selection procedure utilizing principal component analysis (PCA) are used in this model, which

Produced 26 essential elements. In this investigation, a total of 40 input options areutilized. Due to confidentiality concerns, it is not allowed from disclosing the specifications and historical background ofthe features.

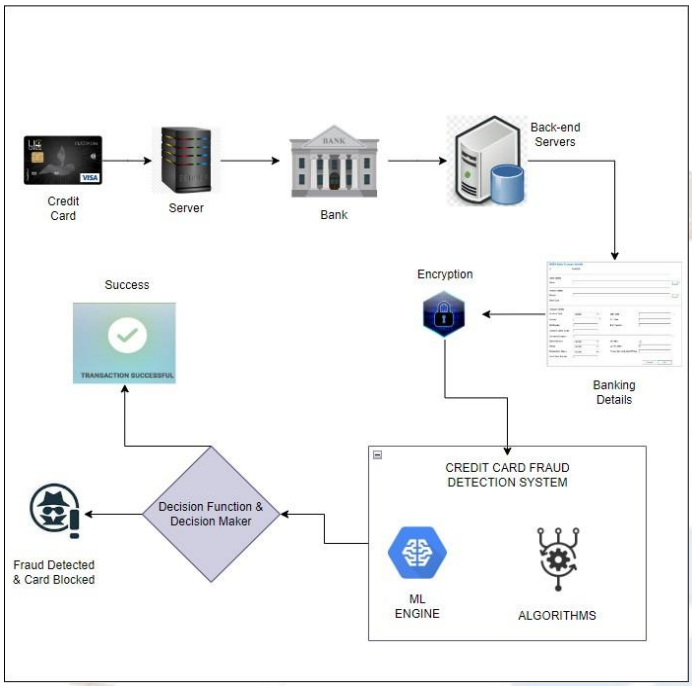

Fig 1. Architecture Diagram of Credit Card FraudDetection System

Logistic Regression

In its essence, a logistic function is used in this statistical model as a binary variable that is dependent to be modeled. This paradigm is mostly applied in situations where a binary classification problem could arise. On classes with linear separability, it performs well. The odd ratio can be used todefine the logic function. It is the likelihood that an occurrence will occur.

Odds Ratio: (l - p) p / p

Where p is the likelihood of a favorable event.

The odds ratio's logarithm is the logit function. It translates input within the [0, 1] range into values inside the real-number range. The following definition applies to the logit function:

$$\text{Logit}(P)=\log {}^{p}/_{1-p}$$

The sigmoid in this model,

$$\Phi(z)={}^{1}/_{1-e}{}^{-z}$$

Logistic regression is one of the most prevalent methods of classification in machine learning. The logistic regression model can be accustomed to define relationships between a continuous, binary, or categorical predictor and other predictors. Variables with a binomial correlation are feasible. Based on a few predictors, we create forecasts about whether or not something will happen. We calculate the likelihood that a given collection of predictors falls into each group.

Both classification and regression can be accomplished using logistic regression, however classification is its primary application. A binary from one of the classesis the output. It is employed to forecast output using dependent variables. With this approach, binary classification to two values is simple. 0 or 1

$$P=1/1+e^{-(a0+a1x1+a2x2+\ldots\ldots+anxn)}$$

The logic behind the logistic regression is described in Eq. 2.In the equation above, a0, a1..an. are coefficients, x1, x2..... xn are independent variables, and p is the result.

B.Random Forest Algorithm

With an abundance of decision trees for different subsets of the dataset, Random Forest is an approach to classification that enhances the overall precision by averaging the correctness of all the decision trees. As the number of decision trees grows, so doesthe random forest's effectiveness. The random forest has characteristics similar to the decision tree, but it incorporates many more from which a better result can be anticipated.

This model is essentially a combined classifier that uses a large number of decision tree classifiers, or an ensemble classifier. The major goal of using many trees is to be able to sufficiently train them so that each tree can contribute in the form of a model. The output is mixed using a majority decision after tree establishment. Itemploys several decision trees, each of which is determined by a unique dataset that is evenly distributed throughout the tree.
This particular model's capacity to account for inaccuracies can effectively balance a population of a class using distinct sets of data. It can be utilized to address problems with classification and regression.

Random Forest is one of the well-liked supervised learning strategies. This can be applied to both regression and classification. But this method's primary use is to categorize challenges, information gained, and it categorizes fraud situations from Non-fraud instances.

RANDOM FOREST ALGORITHM:

To produce classifiers for C:

Randomly choose the training data D with replacement for i=1 to c to produce Di

Make a root node N with Di and cell in it. Create

Tree(N)

Finish for Majority Vote Tree Build(N)

Pick x% at random from N's possible separating characteristics.

Choose the F-features with the most information.

Gain from additional splitting

Gain (T, X) = Entropy (T), minus Entropy (T,X)

Now we create f child nodes in order to calculate the entropy.

Do for i=1 to f.

Put the contents of N into Di Call. Create Tree (Ni)

End for

End

C.K-Nearest Neighbors

A similarity metric, such as the Euclidean, is used by the k-nearest neighbor method of instance-based learning to classify objects. The first two distance indicators perform well with constant variables, but the third distance metric works well for categorical variables.

In this study, the Euclidean distance metric is used to train a k-nearest-neighbor classifier. Between two input vectors (Xi, Xj) for the Euclidean distance (Dij) is givenby:

$$D_{ij} = \sqrt{\sum_{k=1}^{n} \left(X_{ik} - X_{jk}\right)^2} \quad k=1,2,\ldots,n$$

Every point of data in the dataset has its Euclidean distance across an input position and the present position calculated. After these distances are sorted in ascending order,The predominant category is selected by the classifier among these items and offers it as the input point's categorization. The parameter for k has been modified for k = 1, 3, 5, 7, 9, and 11, and k = 3 exhibited the best performance. Consequently, the classifier uses k = 3 as a value.

One of the most straightforward but efficient models is the k-Nearest Neighbor algorithm. The label of the class of the nearest training data pieces in this model determines the test data for the class name. The Euclidean distance [4][16] has been employed to compare two elements' similarity. It is referred to as an instance learning model or an inefficient model. The number of nearest neighbors that must be taken into account is determined by the value of "k."

The value of "k" should be selected appropriately. A suitable distance metric is also necessary. The "Minkowski" distance may occasionally be employed. It is a broadening of the Manhattan and Euclidean distances. It has the following mathematicalrepresentation:

4. Result

This work develops three classification algorithms based on K-Nearest Neighbor algorithm, logistic regression algorithm, and random forest algorithm. The remaining 30% of the dataset is used to train these models, with the remaining 30% designated for testing and verification of these models. Precision, recall, accuracy, f1-score, and ROC curve are used to assess the effectiveness of all three classifiers.

A.Evaluation Criteria

To compare different strategies, we must examine characteristics like accuracy, recall, F1-score and precision. The best methodology to determine credit card fraud will then be applied.

B.Result Analysis

For all three techniques, the ROC curve and the confusion matrix are presented. The dataset produces various results when used with various methods.

The outcomes of applying the dataset for all 3 models are as follows:

Factors	Logistic Regression	Random Forest Algorithm	K-Nearest Neighbors
Accuracy	85.71%	91.55%	97.99%
Precision	87.00%	93.10%	97.00%
Recall	54.00%	72.69%	81.48%
f1-score	67.05%	75.00%	85.00%

TABLE I. OUTCOME FOR 3 MODELS

Logistic Regression

Output

	precision	recall	f1-score	support
Fraudulent	0.87	0.54	0.67	98
Non_Fraudulent	1.00	1.00	1.00	56864
accuracy			1.00	56962
macro avg	0.93	0.77	0.83	56962
weighted avg	1.00	1.00	1.00	56962

0.9799370930139079

Fig 2. Output for Logistic Regression

ROC Curve

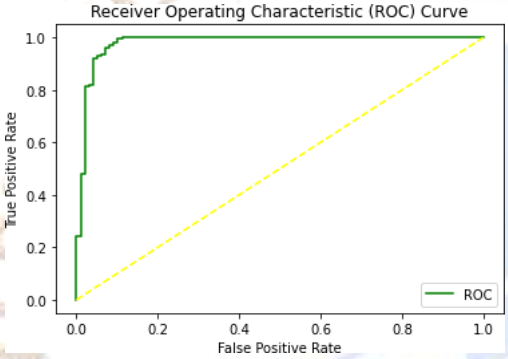

Fig 3. ROC Curve for Logistic Regression

K-Nearest Neighbors

Output

	precision	recall	f1-score	support
Fraudulent	0.96	0.77	0.85	98
Non_Fraudulent	1.00	1.00	1.00	56864
accuracy			1.00	56962
macro avg	0.98	0.88	0.93	56962
weighted avg	1.00	1.00	1.00	56962

0.938721047999954

Fig 4. Output for K-Nearest Neighbors

ROC Curve

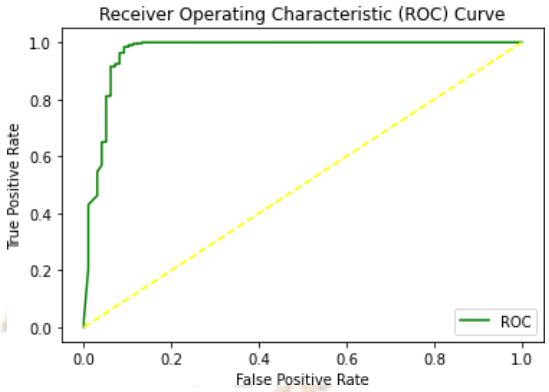

Fig 5. ROC Curve for K-Nearest Neighbors

Random Forest Algorithm

Output

	precision	recall	f1-score	support
Fraudulent	0.97	0.70	0.82	98
Non_Fraudulent	1.00	1.00	1.00	56864
accuracy			1.00	56962
macro avg	0.99	0.85	0.91	56962
weighted avg	1.00	1.00	1.00	56962

0.9672799691063821

Fig 6. Output for Random Forest Algorithm

ROC Curve

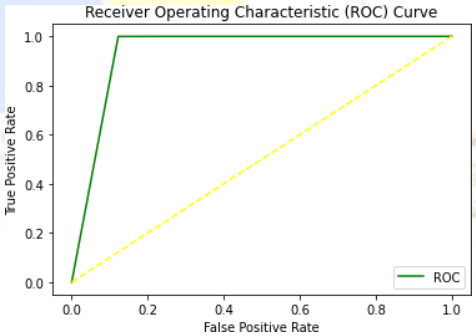

Fig 7. ROC Curve for Random Forest Algorithm

5. Conclusion

Although there are multiple methods to discovering fraud, we cannot guarantee that this particular algorithm fully uncovers the deception. Our analysis leads us to the conclusion that the accuracy of all three algorithms are equivalent. The K-Nearest Neighbors outperforms the Logistic Regression and Random Forest algorithm in terms of precision, recall, and the F1-score. Therefore, we draw the conclusion that the K-Nearest Neighbor algorithm detects credit card fraud better than the other two methods. The future scope of this paper is to bring 100% accuracy and precision by using more efficient upcoming algorithms.

7. Reference

[1] Thennakoon, A., Bhagyani, C., Premadasa, S., Mihiranga, S. and Kuruwitaarachchi, N., 2019, January. Real-time credit card fraud detection using machine learning. In 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 488-493). IEEE.

[2] Sailusha, R., Gnaneswar, V., Ramesh, R. and Rao, G.R., 2020, May. Credit card fraud detection using machine learning. In 2020 4th international conference on intelligent computing and control systems (ICICCS) (pp. 1264-1270). IEEE.

[3] Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M. and Anderla, A., 2019, March. Credit card fraud detection-machine learning methods. In 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-5). IEEE.

[4] Awoyemi, J.O., Adetunmbi, A.O. and Oluwadare, S.A., 2017, October. Credit card fraud detection using machine learning techniques: A comparative analysis. In 2017 international conference on computing networking and informatics (ICCNI) (pp. 1-9). IEEE.

[5] Shukur, Hamzah Ali, and Sefer Kurnaz. "Credit card fraud detection using machine learning methodology." International Journal of Computer Science and Mobile Computing 8.3 (2019): 257-260.

[6] Maniraj, S. P., et al. "Credit card fraud detection using machine learning and data science." International Journal of Engineering Research 8.9 (2019): 110-115.

[7] Khatri, Samidha, Aishwarya Arora, and Arun Prakash Agrawal. "Supervised machine learning algorithms for credit cardfraud detection: a comparison." 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2020.

[8] Popat, Rimpal R., and Jayesh Chaudhary. "A survey on credit card fraud detection using machine learning." 2018 2nd international conference on trends in electronics and informatics (ICOEI). IEEE, 2018.

[9] Tanouz, D., Subramanian, R.R., Eswar, D., Reddy, G.P., Kumar, A.R. and Praneeth, C.V., 2021, May. Credit card fraud detection using machine learning. In 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 967-972). IEEE.

[10] Sudha.C, Nirmal Raj.T, Credit Card Fraud Detection using K-Nearest Neighbor Algorithm, IPASJ International Journal of Computer Science (IIJCS), Volume 5, Issue11, November 2017