**Article title:** Credit Card Fraud Detection Techniques: A Survey
**Authors:** Muhammad Hazeel Ahmed[1]
**Affiliations:** SST, University of Management and Technology, University of Management & Technology (UMT) C-II Block C 2 Phase 1 Johar Town, Lahore, Punjab 54770, Pakistan[1]
**Orcid ids:** 0000-0003-3869-8558[1]
**Contact e-mail:** hazeelbutt2016@gmail.com

**Links to data:** https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud
**DOI:** 10.14293/S2199-1006.1.SOR-.PPFI7P0.v1
**Preprint first posted online:** 06 July 2022
**Keywords:** Fraud detection, Random forest, Fraudulent behavior detection.

# Credit Card Fraud Detection Techniques: A Survey

Muhammad Hazeel Ahmed

June 20, 2022

**Correspondence**
University of Management and Technology
Email:
hazeelbutt2016@gmail.com

**ABSTRACT**

Fraud events are increasing on day to day bases. Credit card fraud is the most frequent fraud that is making a big financial loss on a global level. Researchers have implemented many machine learning algorithms to detect the credit card frauds. This research has briefly described several algorithms and compares the performance of Random Forest, Naïve Bayes, K-Nearest Neighbor, Logistic Regression and Multilayer Perceptron. Algorithms are also used to classify the real transactions or fraudulent transactions. These datasets are compared on the basis of accuracy, precision, recall & false positive rate. Comparison results show that Random Forest performs best in credit card fraud detection dataset among others. Research shows that any ML algorithm can be used to demonstrate the classification of fraud detection.

**KEYWORDS**

Fraud detection, Credit card fraud detection, Random forest, Machine learning, fraudulent behavior detection.

## LAYOUT DESCRIPTION

- Fraud is a criminal activity of a human being, which might be illegal act of money transferring from one's account without notifying. In [1], the author explains Fraud as to misuse of someone's money or assets for one's own advancement.
- Now a days, in transactional frauds security are the major issue. According to the researchers different organizations of the world suffer 5% to 10% lose in their revenues due to fraudulent activities.
- Fraud Detection is a technique to protect a system from fraud. It can also be used to enhance the security of the system in order prevent it from a fraudulent activity.

### What this research adds on to the fraud detection?

1. We have categorized and explored in a positive sense to discuss a detailed knowledge about the scenarios in which fraud detection techniques are used by researchers.
2. We also highlight different fraud detection techniques, some issues and challenges related to it. Further, we discusses analysis of different machine learning algorithms in credit card fraud detection.
3. Future optimization and enhancement in fraud detection system.

We summarize different techniques and algorithms of fraud detection from the research of 2016 till 2021. It also provides positive knowledge about transactional fraud that are occurring currently though credit cards these days.

## I. INTRODUCTION

In today's world E-commerce systems are trending and every single person is aware of it. Now the question arises with the enhancement in electronic commerce systems, frauds are also increased on the day to day bases because these e-commerce system has security issues regarding the privacy of customers' credit cards. These E-commerce websites have two types of users, i.e. Authorized users & Scammers.

Advancement in the system makes users aware of online transactions through credit cards, which is the most easiest and popular way of money transferring on a commercial scale [2]. There can be two types of credit card frauds i.e. fraudster can make an unknown identity to make an online fraud transaction or fraudster can use the stolen or lost credit card to transact money or get cash.

However, transactional fraud is increasing rapidly and every year it is a great loss of money globally [3]. In today's world a person does not need a physical credit card to make any transaction online, but just the information on a credit card is needed to process the transaction. Credit card has very sensitive information, i.e. Card Number, CVV code and expiry date. If a fraudster can have these three private details, then he can make a large amount of purchase.

To overcome the transactional frauds organizations uses different machine learning algorithms for fraud detection. We have analyzed some algorithms of machine learning, i.e. Random Forest, which is used as a classifier to classify normal & fraudulent behaviors by using voting of base classifiers in the data set. [4] Naïve Bayes (NB), Logistic Regression (LG), Multilayer Perceptron (MLP) and decision trees [5] to identify that which algorithm fits best in credit card fraud detection.

This paper also highlights the consequences of financial literacy [6], which is mainly focused on the knowledge of finance and financial decisions in order to carefully manage the money and transactions of credit card. So, by combining the data results of the above analysis from all techniques and algorithms, this paper positively shows that which algorithm will better perform on the commercial scale. Section 2, defines the literature review and related work of previous researchers. Section 3, defines the research methods and techniques which are adopted. Section 4, brings the results with outcomes of doing the research. At last section 5 contains the concluding part of this research paper.

## II. LITERATURE REVIEW

In history, fraud detection has always been a hot topic towards the research. Different researchers have done research in many areas of fraud. In [1], the author refers to the techniques of Artificial Intelligence where he explains the working of neural network with the central nerve system of animals, because it is capable to learn and recognize easily. Researchers represented many types of frauds i.e. bankruptcy frauds, theft fraud and credit card frauds. Paper [1], highlights some of the issues & challenges related to a fraud detection system such as Concept Drift, Real Time Detection, and Skewed Distribution & Large amount of Data. It also provides an overview of different state-of-art Fraud Detection System approaches & methods such as artificial neural network (ANN), support vector machines (SVM), decision trees & meta-heuristics etc. They conclude by highlighting some challenges that are improper to model & have weak accuracy.

According to data mining concept of classification fraud detection falls in the bucket of classification problem [2]. As fraud detection works on the algorithm of data mining to classify the credit card transaction as an original or fraudulent one. The author proposed in [2] that Credit Card Fraud Detection is a problem of Data Mining and there

are two major reasons for which credit card fraud detection is becoming more complex & challenging. They also performed a performance test on the bases of comparison on European cardholders having 284,807 transactions by using three techniques K-nearest, Naive Bayes & Logistic Regression. They conclude by showing the effect of hybrid sampling.

There are two types of credit card frauds, one is a fraud from the application in which fraudster gets a card by using false or wrong information & second is a fraud through stealing the card number and password of card owner and make transactions refer to as behavioral fraud [7]. In [8], misuse detection and anomaly detection are two fraud detection techniques used in for the detection of fraudulent transaction. Misuse detection usually identifies known transactions because it is trained on fake transactions, whereas anomaly detection identify novel ones' because it is trained on normal transactions.

Researchers have classified algorithms of machine learning which are helpful for analyzing the results. In [9], researchers combined LR, SVM, GB and RD on European dataset which provide 91% of the overall. In [4] they combined three techniques LR, DT and RF. So, after analyzing and exploring the techniques they have come up with the best performance of RF with 95.5% and DT with 94.3% and LR with 90%. Paper [4], highlights the classification of Credit Card transactions that they are genuine or fraud.

Basically, credit card fraud detection is based on the behavior of card owner that how it uses. Normally, variables are predicted to know the behavior and this selection has a great effect on the performance of fraud detection systems [5]. Analysis of machine learning algorithms and algorithms of Bayesian network empirically performed better to the economic efficiency. In [5], author tells us about two major problems due to which fraud detection is becoming more complex & challenging, i.e. Profile of a fraudster / normal behavior change and highly skewed data. They also perform comparison in between different algorithms to measure the best out of it, these methods include quadrant discriminative analysis, pipelining and ensemble learning on CCFD.

## III. METHODOLOGY

We come up with the search space that includes different e-databases to acquire valid papers about fraud detection. We assessed different online research papers and explored the work of previous researchers to provide a positive review that will help future researchers have a better understanding of fraud detection techniques.

### a. DATASET DESCRIPTION

For this research we have used a dataset provided on Kaggle named as Credit Card Fraud Detection, which is available at [10]. It was an unbalanced dataset which contains 284,807 total transactions. This dataset contains different attributes such as Class, Amount and Time etc.

Amendment of Principle Component Analysis (PCA) was performed on this dataset. Out of 284,807 transactions only 492 transactions were recognized as fraudulent and 284,315 transactions were recognized as original ones.

For further production dataset has been preprocessed by using the selector tool provided at [11], which is used to remove the features that do not have collective significance towards the results. It is an unbalanced dataset, so we have seen different data sampling techniques such as ADASYN [12] and SMOTE [13] to implement it on this dataset to get balanced data for the future research.

## b. TECHNIQUES / MODELS

- **RANDOM FOREST:**

  Is a machine learning algorithm. It can be used as classification as well as regression, working of random forest is used to make decision trees. It takes the input of samples from datasets and produce decision trees on every sample. Random forest produces decision trees with some of the results and at last it combines or merge the results of different trees to provide accuracy in prediction. This technique is used to train the dataset in two parts i.e. normal behavior & fraudulent behavior [3].

- **K NEAREST NEIGHBORS:**

  It uses different mathematical operations to produce a better output for input datasets. Three main operations are used, i.e. Manhattan distance, Minkowski distance & Euclidean distance. These three measures of similarity are also discussed in [2]. In this paper Euclidean distance is used to find out the classifiers of nearest neighbors. It is the distance between two state points in space is the length of a line segment between two points.

$$\mathbf{d(p, q)} = \sqrt{\sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{q_i} - \mathbf{p_i})^2}$$

  It is calculated between current & new input for every single data point and results re-establish in ascending order and those were selected which have low distance to the input.

- **MULTILAYER PERCEPTRON:**

  It is a multilayer artificial neural network having 3 different layers, i.e. input, output & hidden layers. Input layer processed the signals of input. Hidden layer is placed between input & output layer and Output layer performs the classification [14]. It uses functions which calculate the weight and adds bias to its output. Its function will be look like as**:**

$$\boldsymbol{f(x) = f(WxT + b)}$$

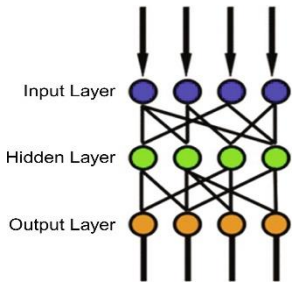


**Fig 1:** Pictorial representation of MLP

- **NAIVES BAYES:**

  It is a decision maker, which make decisions on the bases of high probability [2]. It is one of the algorithm in which attribute dependencies are not applicable & its probability always uses the values that are known in order to produce the estimated probability of unknowns. In this paper the Naïve Bayes classifier is used as fraudulent or non-fraudulent.

- **LOGISTIC REGRESSION:**

  It is a machine learning algorithm popularly known for its working, while calculating probabilities it make dependent variable binary. It will predict that something will happen or not. It also uses a function that is called the sigmoid to find the best predictions of the results. Values of a function will be treated as 0 if it is less than 0.5 and considered 1 if it is greater than 0.5.

## IV. RESULTS & DISCUSSIONS

We have discussed five techniques for the detection of fraudulent transactions. We use a concept of classification evaluation described in [15], which include Confusion matrix having 4 different sets of values, i.e. True Positive (TP), False Positive (FP), False Negative (FN) and True Negative (TN). For comparison, we have considered its Recall, Precision, Accuracy and False Positive Rate. We also use ROC Curve & Precision Recall Curve for some pictorial representation of the dataset.

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

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

$$\text{Accuracy} = \frac{TP + TN}{N}$$

$$\text{False Positive Rate} = \frac{FP}{FP + TN}$$

We have discussed models of classification above and used them on the dataset. In below given tables Precision, Recall and False Positive Rate basically shows the fraudulent transaction in the dataset.

**Table 1:** Represent Results of Classification Evaluation

| CLASSIFIERS | RECALL | PRECISION | FPR | ACCURACY |
|---|---|---|---|---|
| Random Forest | 0.89 | 0.31 | 0.46 | 99.7% |
| K-Nearest Neighbor | 0.55 | 0.02 | 0.03 | 94.4% |
| Multilayer Perceptron | 0.90 | 0.08 | 0.15 | 98.4% |
| Naïve Bayes | 0.85 | 0.26 | 0.40 | 99.6% |
| Logitic Regression | 0.91 | 0.07 | 0.14 | 98.2% |

As we can see that Random Forest gives highest accuracy and K-Nearest Neighbor gives lowest accuracy among these models of classification. In our experiment Random forest performs better than other four models, but on the other side for any other dataset it might be possible that some other model will work better than Random forest. We have created a bar chart of comparison on the basis of accuracy and we have constructed ROC_Curve of Random Forest Classifier to demonstrate the pictorial graph.

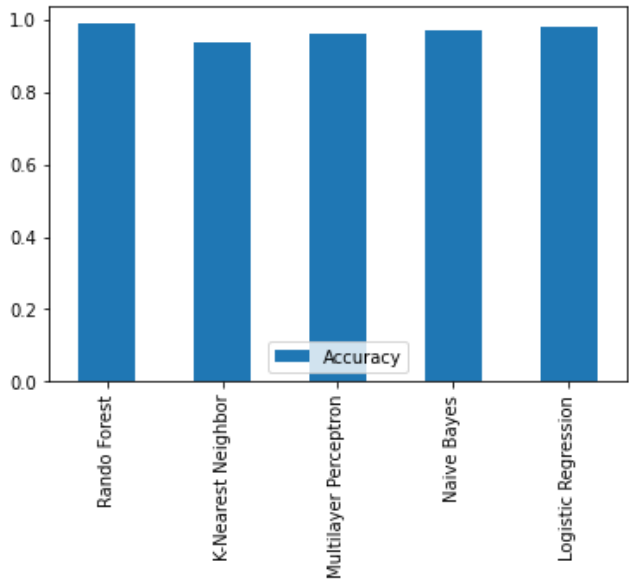


**Fig 2:** Comparison

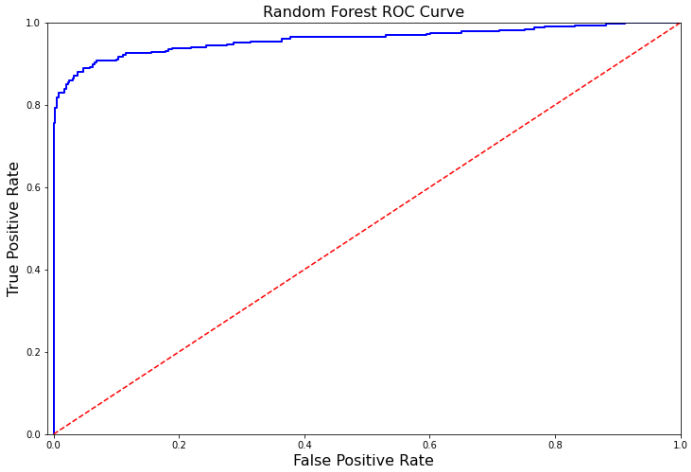


**Fig 3:** ROC Curve

## V. CONCLUSION

Credit Card fraud has been increasing in recent years, these fraudulent transactions, from credit cards are causing serious money loss at global level to the people using bank accounts for the safety of their money. Many Fraud detection systems have been proposed technically to prevent the fraudulent activities. Our findings are based on the original dataset of credit card fraud detection. The main goal of this paper is to make a comparison between different techniques & models of ML. This research shows that Random Forest gives better results among other algorithms of ML. We have extracted the results of classification evaluation in tabular forms to provide better understanding. Furthermore, we have created a bar chart of comparison on the bases of accuracy to give a pictorial representation of the results. All our findings are actual and reality-based. Future researchers should make a comparison with other models in order to provide a wider view of the fraud detection techniques. We hope that this paper will help the researchers have positive minds towards Credit Card Fraud Detection.

## VI. ACKNOWLEDGEMENT

I'm thankful to Dr. Sheraz Naseer, who is having a vast knowledge of Data Science for providing their valuable time and guidance. I'm pleased with the support of Dr. Sheraz Naseer because he helped me to do quality research and how to write a quality paper by using different tools and techniques. This paper is written by Muhammad Hazeel Ahmed from the University of Management and Technology.

## VII. REFERENCES

[1] Aisha Abdallah, M. A. Maarof, and A. Zainal, "Fraud detection system: A survey," *J. Netw. Comput. Appl.*, vol. 68, pp. 90–113, Jun. 2016, doi: 10.1016/j.jnca.2016.04.007.

[2] John O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis," in *2017 International Conference on Computing Networking and Informatics (ICCNI)*, Oct. 2017, pp. 1–9. doi: 10.1109/ICCNI.2017.8123782.

[3] Shiyang Xuan, G. Liu, Z. Li, L. Zheng, S. Wang, and C. Jiang, "Random forest for credit card fraud detection," in *2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC)*, Mar. 2018, pp. 1–

6. doi: 10.1109/ICNSC.2018.8361343.
[4] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, and A. Anderla, “Credit Card Fraud Detection - Machine Learning methods,” in *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*, East Sarajevo, Bosnia and Herzegovina, Mar. 2019, pp. 1–5. doi: 10.1109/INFOTEH.2019.8717766.
[5] Siddhant Bagga, A. Goyal, N. Gupta, and A. Goyal, “Credit Card Fraud Detection using Pipeling and Ensemble Learning,” *Procedia Comput. Sci.*, vol. 173, pp. 104–112, Jan. 2020, doi: 10.1016/j.procs.2020.06.014.
[6] A. Saleh Hussein, R. Salah Khairy, S. M. Mohamed Najeeb, and H. Th. S. Alrikabi, “Credit Card Fraud Detection Using Fuzzy Rough Nearest Neighbor and Sequential Minimal Optimization with Logistic Regression,” *Int. J. Interact. Mob. Technol. IJIM*, vol. 15, no. 05, p. 24, Mar. 2021, doi: 10.3991/ijim.v15i05.17173.
[7] “[PDF] Unsupervised Profiling Methods for Fraud Detection | Semantic Scholar.” https://www.semanticscholar.org/paper/Unsupervised-Profiling-Methods-for-Fraud-Detection-Bolton-Hand/5b640c367ae9cc4bd072006b05a3ed7c2d5f496d (accessed Jun. 11, 2022).
[8] A. Kundu, S. Sural, and A. Majumdar, “Two-Stage Credit Card Fraud Detection Using Sequence Alignment,” 2006. doi: 10.1007/11961635_18.
[9] “Amusan et al. - 2021 - Credit Card Fraud Detection on Skewed Data using M.pdf.”
[10] “Credit Card Fraud Detection.” https://www.kaggle.com/mlg-ulb/creditcardfraud (accessed Jun. 11, 2022).
[11] W. Koehrsen, *Feature Selector: Simple Feature Selection in Python*. 2022. Accessed: Jun. 11, 2022. [Online]. Available: https://github.com/WillKoehrsen/feature-selector
[12] “ADASYN: Adaptive synthetic sampling approach for imbalanced learning | IEEE Conference Publication | IEEE Xplore.” https://ieeexplore.ieee.org/document/4633969 (accessed Jun. 11, 2022).
[13] “SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary | Request PDF.” https://www.researchgate.net/publication/325789071_SMOTE_for_Learning_from_Imbalanced_Data_Progress_and_Challenges_Marking_the_15-year_Anniversary (accessed Jun. 11, 2022).
[14] “Multilayer Perceptron - an overview | ScienceDirect Topics.” https://www.sciencedirect.com/topics/computer-science/multilayer-perceptron (accessed Jun. 12, 2022).
[15] “Confusion Matrix - an overview | ScienceDirect Topics.” https://www.sciencedirect.com/topics/engineering/confusion-matrix (accessed Jun. 12, 2022).