International Journal of Social Science Research (IJSSR)
Volume- 3, Issue- 3 | May - June 2026 ISSN: 3048-9490

A Comparative Study of Machine Learning Methods for Credit Card Fraud Detection

Mithun Das[2*], Tapas Roy[1], Dr. Rajesh Das[3]
[1]PhD Scholar, Department of Library and Information Science, The University of Burdwan
https://orcid.org/0009-0007-6482-6425
[2]PhD Scholar, Department of Library and Information Science, The University of Burdwan
https://orcid.org/0009-0003-0610-5057
[3]Assistant Professor, Dept. of Library and Information Science, The University of Burdwan
http://orcid.org/0000-0001-5349-6589

Abstract

The problem of credit card fraud causes significant financial losses to businesses and consumers and prompts studies on effective ways of detecting fraud. Here we examine a publicly available dataset of credit card transactions (a total of 10,000) (151 fraudulent, 9,849 legitimate) and create supervised learning models to categorize credit card transactions. We train classifiers of logistic regression, random forest, and XGBoost, and the class imbalance will be addressed through the use of class-weighting. Performance is also assessed using accuracy, recall, F1-score, and ROC AUC (Receiver Operating Characteristics Area under the Curve) as sole accuracy will not suffice in the unbalanced data scenario. Evidence indicates that logistic regression achieves high recall (91) but low precision (23%), which is a high number of false alarms. Compared to the above, random forest is balanced (precision 100, recall 58, F1 =0.73), and the XGBoost is almost perfect in discrimination (precision 100, recall 97.8, F1 =0.99, ROC AUC =0.999). The results are consistent with available literature, which indicates that ensemble techniques (random forests and boosting) are more effective than simple models in detecting fraud. We speculate about the implications of this trade-off on a real-world deployment, as well as providing future work directions.

Keywords: Machine learning, Credit card, Credit card fraud, XGBoost, Logistic regression, Financial losses.

Introduction

Credit card fraud has remained a major menace to financial institutions across the globe with loss amounting to hundreds of millions annually. With the spread of digital transactions, attackers are using more advanced methods, which makes it essential to detect them in real-time. Machine learning has established itself as a part and parcel of fraud detection systems, as it is capable of automatically determining complicated patterns that reveal fraudulent or legitimate transactions. This is because past studies have proven methods like logistic

*Corresponding Author Email: 02mithun02@gmail.com
Published: 09 May 2026
DOI: https://doi.org/10.70558/IJSSR.2026.v3.i3.301059

regression, decision trees, support vector machines, and ensemble methods (random forests, gradient boosting) to be effective in the field. Nevertheless, credit card data are highly skewed (fraud is not common), and overall accuracy is not as useful as a metric. This paper involves a 10K data on transactions (obtained through a Kaggle repository) to construct and compare supervised classifiers, with special attention to precision, recall, and AUC to overcome the imbalance. We would like to determine the most appropriate modeling technique that would optimally trade between the false alarms and the detection rate on this data.

Literature review:

Dornadula and Geetha (2019) have shown that credit card fraud detection is one of the critical challenges because the e-commerce and real-time digital payment systems are rapidly developing. The conventional batch-based machine learning methods have proved to be weak in case of streaming transaction data especially in dealing with changing customer behavior and concept drift. To mitigate this, scholars have considered the idea of sliding windows to model transaction trends in time and dynamically adjust models as time goes by. Grouping cardholders according to transactional characteristics including spending patterns has been observed to boost the accuracy of detecting fraud due to the ability to learn group specific models. Moreover, comparative studies on various classifiers show that ensemble and adaptive learning techniques are superior to the use of static models in warning of fraudulent activities in the dynamic transaction settings.

Varmedja et al. (2019) have conducted a study that considered credit card fraud as the abuse of information that belongs to the cardholder because of the physical loss of the card or the unauthorized access to sensitive financial materials. The use of machine learning methods has become common to distinguish between a fraudulent and a legitimate transaction, and the following algorithms have proven to be effective in classification: Logistic Regression, random forest, Naive Bayes, and Multilayer Perceptron. Since credit card transaction datasets are highly imbalanced, oversampling methods like SMOTE (Synthetic Minority Over-sampling Technique) are used by researchers to add representation to minority classes. It has been demonstrated that feature selection and proper partitioning of the train and the test data can improve the generalization and predictive accuracy of the models. The results of experimental studies conducted by several researchers point to the fact that these models have high detection rates and can be generalized to detect other types of financial anomalies.

Alarfaj et al. (2022) have written a paper on fast increase in online credit card transaction has contributed to the high number of fraud transactions, which has caused significant financial losses by both financial institutions and consumers. The research findings produced by earlier studies have used different machine learning methods, such as Decision Trees, Random Forests, Support Vector Machines, Logistic Regression, Extreme Learning Machines, and XGBoost, to solve the problem of fraud detection. Nevertheless, the methods are frequently challenged by the difficulties in the form of extreme class imbalance, fraud patterns development, and false alarms. Recent research points to the possibility of deep learning models to put into practice complex and non-linear transaction process more efficiently than conventional techniques. In particular, convolutional Neural Network-based architectures have been found to be more effective in cases where the depth and training parameters are optimised.

Different benchmark European credit card datasets consistently depict that deep learning methods have better accuracy, precision, F1-score, and AUC, which qualifies them as good solutions in actual fraud detection systems.

Itoo, Meenakshi, and Singh (2020) have conducted a research work on the rising rate of financial fraud has become one of the most pressing issues because it has a negative economic effect on financial institutions and regulation. Previous researches reveal that the quick development of credit card transactions due to development of internet technology has also exposed chances of fraudulent transactions. The fact that credit card datasets have highly imbalanced data has always been a significant concern on researchers because of the negative impacts on the prediction capabilities of fraud detection models. To address this problem, random under-sampling and over-sampling strategies of data resampling have been extensively used to enhance the representation of the minority classes. Multiple machine learning methods have been studied and compared with extensive performance indexes such as Logistic Regression, Naivete Bayes, and K- nearest neighbours. Empirical evidence indicates that, with proper data balancing methods, logistic regression tends to do better with fraudulent transactions than other conventional classifiers.

Malik et al. (2022) have highlighted a report on the rise in the use of machine learning methods by financial institutions to solve the ever-increasing issue of credit card fraud. The existing literature has investigated both individual and group learning, which has proved that hybrid and group-based frameworks have a tendency of performing better than individual classifiers. However, despite these developments, there is little research that has comparatively studied the efficacy of various hybrid model combinations using the same data. Recent discoveries indicate that the boosting-based hybrids, including the AdaBoost plus gradient-boosted tree models may help to achieve a significant increase in detection accuracy and classification error. The effectiveness of hybrid methods is however very dataset-specific and not every hybridization results in beneficial improvements. Therefore, more empirical research is needed to explore various hybrid models, feature engineering methods, and preprocessing approach to determine the strongest fraud detection models.

Statement of the Problem

Although the digital payment system has made tremendous improvements, credit card fraud still occurs as a very expensive issue among the financial institutions across the globe. The growing amount and rate of online transactions coupled with advanced techniques in committing frauds render the manual or rule-based detection systems ineffective and inefficient. There are two important issues that prevail in this field.

To start with, credit card fraud data are very imbalanced, with the amount of fraud transactions constituting a very small percentage of the total data. This unbalance leads to the conventional machine learning models being biased against the majority (legitimate) class and hence the high overall accuracy but poor fraud detection performance, specifically low recall of the minority class. This has resulted in a lot of fraudulent transactions going unnoticed and this has caused loss of money and loss of customer confidence.

Second, machine learning models have a trade-off between the accuracy of fraud detection and false positives. High-fraud detection (recall) models tend to produce many false positives, indicating that there are genuine transactions with a fraud label. This results in unwarranted decline in the transactions, customer dissatisfaction, and higher operational cost to the financial institutions. One of the major research issues is to come up with a strong model that can efficiently balance between fraud detection and a low number of false alarms.

Objectives:

The following are the main goals of this study:

1. To design and test supervised machine learning to identify credit card fraud under extreme class imbalance, and to confirm their effectiveness using suitable evaluation metrics, including precision, recall, F1-score and ROC-AUC instead of accuracy levels, only.
2. To conduct a comparative analysis of the various machine learning algorithms in respect to their capacity to achieve a balance between fraud detection and reduction of false positives with the view of coming up with a model that offers optimal performance in a real-life financial transaction monitoring system.

Methodology

Data Description:

For this research, we download the credit card fraud detection dataset (360.52 kB) CSV format from www.kagggle.com. The characteristics of this dataset are i). Total records 10000. ii). Total features 10. iii). target variables is _fraud. iv). class distribution: highly imbalance (fraud=4-5%). And the feature descriptions of the data set are given in the form of a table.

Table-1 Feature Description:

Sl. no.	Feature Name	Description
1	transaction_id	Unique identifier for each transaction
2	amount	Transaction amount
3	transaction_hour	Hour of transaction (0–23)
4	merchant_category	Type of merchant
5	foreign_transaction	Indicates if transaction is international (0/1)
6	location_mismatch	Billing vs transaction location mismatch (0/1)
7	device_trust_score	Trust score of the device (0–100)
8	velocity_last_24h	Number of transactions in last 24 hours

9	cardholder_age	Age of the cardholder
10	is_fraud	Target variable (0 = Normal, 1 = Fraud)

Preprocessing:

We dropped transaction ID (non-predictive) and one-hot encoded merchant category which was a category. The last feature set comprised of 12 predictors, five binary merchant category indicators and seven numeric variables (amount, hour, foreign flag, mismatch, trust score, velocity, age). In order to manage the uneven distribution of fraud, we conducted a stratified train-test split of 70:30 in both sets such that the proportion of fraud in these sets was exactly equal. All the models were trained using the original class distribution; however, we also applied class-weight modifications to address the problem of imbalance: e.g., classweight=balanced in logistic regression and random forest weights the misclassifications of the infrequent cases of fraud with a larger penalty.

Models:

We used three classifiers that were under supervision. Logistic Regression (LR) offers a baseline that is a simple linear one. Random Forest (RF) is a collection of decision trees with strengths against noise and high-dimensional data covering. We have made use of scikit-learn implementations, where hyperparameters were optimized through grid search (e.g. RF using 100 trees, no depth limit). We also trained an XGBoost classifier that frequently works in lopsided classification. We also use the scale-pos-weight parameter of XGBoost with the negative/positive ratio to further adjust the imbalance. Other sampling (oversampling/ undersampling) was not a part of it.

Evaluation:

As is common practice, we estimated precision (TP/TP+FP), recall (TP/TP+FN), F1-score and ROC-AUC on the held-out test set, which we evaluated using models. These measures represent a trade-off between false alarms (precision) and catching frauds (recall). Confusion matrices were also inspected to know the type of errors (true/false positives/negatives). All the metrics were calculated with the help of the standard functions of scikit-learn to obtain reliability.

Results

Table 2: Summarizes the test-set performance for each model on the fraud (positive) class:

Model	**Precision**	**Recall**	**F1-score**	**ROC AUC**
Logistic Regression	0.230	0.911	0.368	0.989

Random Forest	1.000	0.578	0.732	0.999
XGBoost	1.000	0.978	0.989	0.999

Table 2 presents the performance statistics of the models in the fraud class test set. The logistic regression model gave maximum recall (finds 91.1% of the frauds), which comes at the cost of very low precision (23.0%), indicating some genuine transactions as frauds. There are only 41 correct predictions of actual frauds as compared to the LR confusion matrix (not shown) of false positives of 137. This is in accordance with previous reports: LR is often highly sensitive, however it generates a significant false alarm rate.

The logistic regression model reported many legitimate transactions as fraudulent to maximize recall (91.1% of the frauds) at the cost of very low precision (23.0%). The LR confusion matrix (not shown) showed that only 41 real frauds were anticipated of 137 false positives. This trend is in line with previous reports: LR often generates too many false alarms even when it is very sensitive.

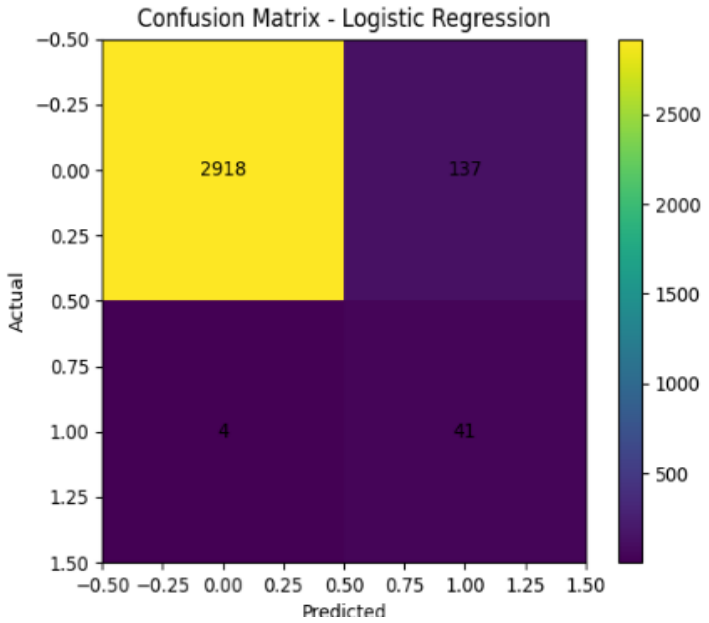

Figure 1: A confusion matrix generated with the model of logistic regression which shows the true positives, true negatives, false positives, and false negatives.

By contrast, the Random Forest model did not detect many scams but it did not give any false positives when we ran it (100 percent precision). Specifically, RF did not erroneously call a legal transaction fraud, correctly marking 26 of 45 cases of fraud (recall 57.8). This was the trade-off (high precision, moderate recall) that made the F1-score (0.732) larger than LR. Although certain frauds may slip through, in most practice situations it is better to have less false alarms.

Random forest:

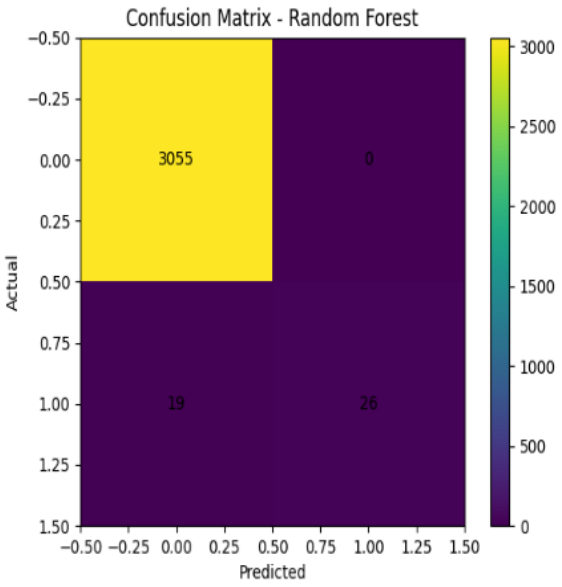

Figure 2: A confusion matrix showing true positives, true negatives and misclassifications in the classification of the random forest model.

The XGBoost model that made the best combination overall was with F1-score of 0.99 and ROC AUC of 0.999: it correctly classified 44 of the 45 frauds (97.8% recall) and did not misclassify any in this test set (100% precision). As a matter of fact, XGBoost separated the classes in this data rather fully. Other reports have indicated the same trend: gradient boosting has better performance than the classical models. We note that these results might be even worse in the real-world on the larger datasets; these astronomical high numbers might be explained at least in part by the small size of the sample and high quality in the available features.

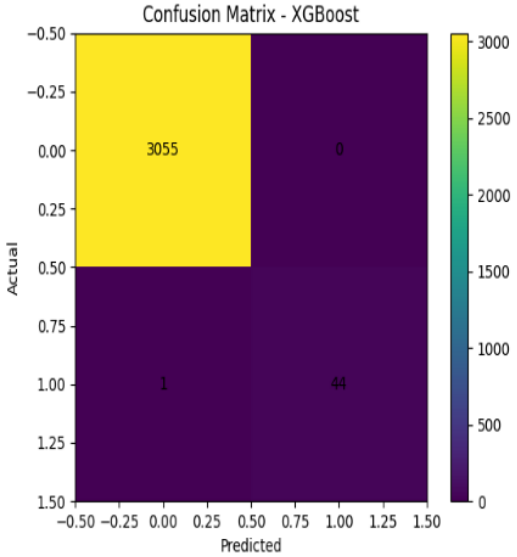

Figure 3: XGBoost Model Confusion Matrix of Credit Card Fraud Detection.

In general, the results are consistent with a common thread of the literature: more complex ensemble methods (RF, XGB) are more accurate and generally effective, and more basic linear models (LR) are more likely to lean towards a false alarm to detect fraud. In this case, Random Forest did a great job to find the balance between specificity and sensitivity and XGBoost did it even better.

Discussion

We have tested that model behavior is largely affected by a class imbalance. Naive accuracy is misleading because the fraud rate is very low (1.5% of the transaction) and is consistent with the best practices. Large recall (detecting most frauds) is often prioritized in receiving fraud detection work, and a large amount of false positives will negatively affect customer experience and increase investigative costs. The logistic regression model was the best characterizing this tension, as it revealed nearly all frauds (recall 91%) but only 23% precise to produce a number of false alarms. Models that reached much higher precision and a minor decrease in recall, however, were the Random Forest and XGBoost.

These tendencies favor previous studies. The ensemble models proved to have the best precision, F1 score, ROC-AUC, and the highest recall and ROC-AUC, but the lowest precision and F1 score, showing that a high rate of false positives (Shahbahrami et al., 2025). We can always obtain the following: RF and XGB tree-based ensembles achieved better F1 and AUC as compared to LR. It means that non-linear interactions of decision trees, including transaction hour, amount, and user/device characteristics, can be helpful to detect the presence of fraud in the given data set. Actually, the two most important attributes in our model of the Random Forest were the time of the transaction and device trust score, implying that suspicious hours and device abnormalities are a major indicator of fraud (as found in the literature).

It has several limitations. One, it is possible that the results will not be applicable to a larger and more heterogeneous dataset due to the small size of the dataset (10K items) and the lack of its complete coverage of the variety of interactions in the real world. Second, we did not apply more advanced techniques such as deep learning models and synthetic oversampling (SMOTE). Other studies have found that neural architectures can become a little more recall-biased at the cost of interpretability and that making the minority class larger can also raise the model robustness. Finally, every one of our models was based on fixed data; in practice, fraud detection systems must be able to adapt to evolving fraud patterns. Future work might solve temporal drift by retraining models on a regular schedule, including unsupervised anomaly detection of new fraud or use federated learning to maintain user privacy whilst using multi-institution data (recommended by recent reviews).

Conclusion

In this research, a sample of 10,000 credit cards transactions was used to evaluate machine learning in fraud detection. We demonstrate that the accuracy-memory trade-offs of classical models differ tremendously: logistic regression identifies most frauds with a high rate of false positives, and random forests and gradient boosting has a high rate of precision with only a minor decrease in recall. XGBoost was the most successful in tests on average (F1 = 0.989, ROC AUC = 0.999). These findings give credence to the previous studies that highlight the

efficiency of ensemble methods in this task. More information, addition of other features (e.g. device IDs and geolocation) and approaches that consider imbalance would further improve detection in the future. As we show, even a basic supervised pipeline can be used with the appropriate measures to identify credit card fraud.

References:

Dornadula, V. N., & Geetha, S. (2019). Credit Card Fraud Detection using Machine Learning Algorithms. *Procedia Computer Science*, *165*, 631–641. https://doi.org/10.1016/j.procs.2020.01.057

Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M., & Anderla, A. (2019, March). Credit card fraud detection-machine learning methods. In *2019 18th International Symposium Infoteh-Jahorina (Infoteh)* (pp. 1-5). IEEE.DOI: 10.1109/INFOTEH.2019.8717766

Alarfaj, F. K., Malik, I., Khan, H. U., Almusallam, N., Ramzan, M., & Ahmed, M. (2022). Credit card fraud detection using State-of-the-Art machine learning and deep learning algorithms. *IEEE Access*, *10*, 39700–39715. https://doi.org/10.1109/access.2022.3166891

Itoo, F., Meenakshi, N., & Singh, S. (2020). Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. *International Journal of Information Technology*, *13*(4), 1503–1511. https://doi.org/10.1007/s41870-020-00430-y

Malik, E. F., Khaw, K. W., Belaton, B., Wong, W. P., & Chew, X. (2022). Credit card fraud detection using a new hybrid machine learning architecture. *Mathematics*, *10*(9), 1480. https://doi.org/10.3390/math10091480

Sulaiman, R. B., Schetinin, V., & Sant, P. (2022). Review of Machine Learning Approach on Credit Card Fraud Detection. *Human-Centric Intelligent Systems*, *2*(1–2), 55–68. https://doi.org/10.1007/s44230-022-00004-0

Roseline, J. F., Naidu, G., Pandi, V. S., Rajasree, S. a. A., & Mageswari, D. (2022). Autonomous credit card fraud detection using machine learning approach. *Computers & Electrical Engineering*, *102*, 108132. https://doi.org/10.1016/j.compeleceng.2022.108132

Chounchenani, M. D., Shahbahrami, A., Hassanpour, R., & Gaydadjiev, G. (2025). Deep Learning Based Image Aesthetic Quality Assessment- a review. *ACM Computing Surveys*, *57*(7), 1–36. https://doi.org/10.1145/3716820