International Journal of Applied Smart Interdisciplinary Technologies (IJASIT)

Home Page: https://ijasit.org **ISSN: 3139-759X**

https://doi.org/10.68104/ijasit.v1.i3.34

Credit Card Fraud Detection Automation System Using Logistic Regression and XG Boost

Harshwardhansinh Chauhan[1,*], Rocky Upadhyay[2]

[1 2] Drs. Kiran & Pallavi Patel Global University (KPGU), Vadodara, Gujarat, India. *hwkc392@gmail.com, rko75316060@gmail.com*

*** Corresponding author.**
E-mail: *hwkc392@gmail.com*

Article History:
Received Date: 09 August 2026
Revised Date: 27 August 2026
Accepted Date: 05 August 2026
Published Date: 13 September 2026

Keywords:

Minimum three keywords; avoid too general and too specific keywords; CDF Letters

ABSTRACT

Credit card fraud poses a significant challenge to financial institutions, merchants, and customers, resulting in substantial financial losses and reduced consumer confidence. The rapid growth of digital payment systems has increased the volume and complexity of transactions, making automated and reliable fraud detection essential. This study proposes an automated machine learning framework for credit card fraud detection that addresses the severe class imbalance inherent in fraud datasets through Random Under-Sampling and Synthetic Minority Over-sampling Technique (SMOTE). The framework evaluates multiple approaches, including Logistic Regression, Random Forest, XGBoost, and an Autoencoder-based anomaly detection model, to provide a comprehensive comparison of supervised and unsupervised learning techniques. Unlike conventional approaches that primarily rely on accuracy, the proposed framework evaluates model performance using Precision, Recall, F1-score, ROC-AUC, and Precision-Recall AUC (PR-AUC), providing a more meaningful assessment of fraud detection capability. An initial Logistic Regression experiment using data preprocessing, exploratory analysis, statistical validation, and under-sampling achieved an accuracy of 93.91%. The extended framework aims to identify models that effectively balance the detection of fraudulent transactions with the reduction of false alarms. The proposed approach provides a scalable foundation for real-time transaction monitoring and can support adaptive fraud detection systems capable of responding to evolving fraud patterns.

1. Introduction

The rapid growth of digital payment systems, online banking, and e-commerce has significantly increased the volume of credit card transactions worldwide. While electronic payments provide convenience, speed, and accessibility, they have also created new opportunities for fraudulent activities. Credit card fraud occurs when unauthorized transactions are performed using a customer's payment information, resulting in financial losses for customers, merchants, and financial institutions. In addition to direct monetary losses, frequent fraudulent transactions can reduce customer confidence and negatively affect the reputation of financial service providers. Therefore, developing accurate, reliable, and automated fraud detection systems has become an important research problem in the financial technology domain.

Traditional fraud detection systems generally depend on predefined rules and manually designed conditions to identify suspicious transactions. Although rule-based approaches can detect known fraud patterns, they may have limited ability to identify new and continuously evolving fraud strategies. Machine learning (ML) provides an effective alternative because ML algorithms can learn patterns from historical transaction data and identify potentially fraudulent transactions automatically. Techniques such as Logistic Regression, Random Forest, and gradient-boosting methods can be used to classify transactions as legitimate or fraudulent. In addition, unsupervised approaches such as Autoencoders can identify unusual transaction patterns without requiring every transaction to be explicitly labelled.

One of the major challenges in credit card fraud detection is the highly imbalanced nature of transaction datasets. In a typical dataset, legitimate transactions significantly outnumber fraudulent transactions. As a result, a model can achieve high overall accuracy simply by predicting most transactions as legitimate while failing to detect a substantial proportion of actual fraud cases. Therefore, accuracy alone is not an appropriate performance measure for fraud detection. Metrics such as Precision, Recall, F1-score, ROC-AUC, and Precision-Recall AUC (PR-AUC) provide a more informative evaluation by considering the ability of a model to detect fraudulent transactions and control false-positive predictions.

To address the class imbalance problem, sampling techniques such as Random Under-Sampling and Synthetic Minority Over-sampling Technique (SMOTE) can be incorporated into the machine learning pipeline. Random Under-Sampling reduces the number of majority-class transactions, while SMOTE generates synthetic examples of the minority fraud class to improve its representation during model training. Combining these techniques with multiple machine learning algorithms enables a systematic comparison of different fraud detection strategies. In this study, Logistic Regression, Random Forest, XGBoost, and an Autoencoder-based anomaly detection approach are investigated within an automated fraud detection framework.

The proposed framework extends an initial experimental approach that included data cleaning, exploratory data analysis, statistical analysis, under-sampling, train-test data splitting, and Logistic Regression, which achieved an accuracy of 93.91%. The extended framework moves beyond accuracy-based evaluation by comparing different models using multiple performance measures, with particular emphasis on fraud detection capability and the trade-off between detecting genuine fraudulent transactions and generating false alarms. The overall objective is to develop a robust and scalable framework that can support automated and real-time credit card fraud detection while providing a foundation for future systems capable of adapting to changing and emerging fraud patterns.

2. Related Work

Credit card fraud detection has been extensively studied using statistical methods and machine learning techniques because of the increasing volume of digital transactions and the highly imbalanced nature of fraud datasets. Early approaches mainly relied on rule-based systems and traditional statistical methods, but their ability to identify previously unseen fraud patterns was limited. Machine learning algorithms such as Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, and Artificial Neural Networks have subsequently been applied to improve automated fraud detection. Among these methods, ensemble techniques such as Random Forest and gradient-boosting algorithms have demonstrated strong classification performance because they can capture complex relationships between transaction features. However, the effectiveness of supervised learning models is strongly influenced by the availability and distribution of labelled fraudulent transactions.

Class imbalance remains one of the most important challenges identified in previous research. Since fraudulent transactions generally represent only a small proportion of total transactions, models trained directly on imbalanced datasets may achieve high accuracy while performing poorly in detecting actual fraud cases. To address this issue, researchers have investigated sampling techniques such as Random Under-Sampling and SMOTE, along with cost-sensitive and anomaly-detection approaches. Autoencoders have also attracted attention because they can learn representations of normal transaction behaviour and identify transactions that deviate significantly from learned patterns. Recent studies increasingly recommend evaluating fraud detection models using Precision, Recall, F1-score, ROC-AUC, and PR-AUC rather than relying only on accuracy. Building on these findings, the present study provides a comparative framework incorporating Logistic Regression, Random Forest, XGBoost, and Autoencoder-based anomaly detection with sampling-based class imbalance handling.

Table 1

Comparison of Existing Approaches

Study/Approach	Technique Used	Imbalance Handling	Evaluation Metrics	Key Finding
Traditional Fraud Detection	Rule-based and statistical methods	Rule-based thresholds	Accuracy, Detection Rate	Effective for known patterns but limited for emerging fraud
Logistic Regression	Supervised classification	Under-Sampling	Accuracy, Precision, Recall	Simple and interpretable baseline model
Random Forest	Ensemble learning	Under-Sampling / SMOTE	Precision, Recall, F1-score, ROC-AUC	Handles nonlinear relationships and provides robust classification
XGBoost	Gradient-boosting ensemble	SMOTE / sampling	Precision, Recall, F1-score, ROC-AUC, PR-AUC	Strong predictive capability for complex transaction patterns
SMOTE-Based Approaches	Synthetic minority generation	SMOTE	Precision, Recall, F1-score	Improves representation of the minority fraud class
Autoencoder	Unsupervised anomaly detection	Normal transaction learning	Reconstruction Error, Precision, Recall	Detects unusual transactions without relying entirely on fraud labels
Proposed Framework	LR, RF, XGBoost, Autoencoder	Under-Sampling + SMOTE	Precision, Recall, F1-score, ROC-AUC, PR-AUC	Provides a comprehensive comparison focused on fraud detection and false-alarm reduction

3. Research Gap

The existing literature demonstrates substantial progress in machine learning-based credit card fraud detection; however, several important research gaps remain. First, many implementations focus primarily on a single classification algorithm, such as Logistic Regression, which limits the ability to determine whether the selected model is optimal for fraud detection. Second, evaluation is often centered on overall accuracy, with the reported 93.91% accuracy providing limited insight into the model's ability to correctly identify fraudulent transactions. Since fraud datasets are highly imbalanced, a model may achieve high accuracy while failing to detect a significant number of actual fraud cases. Therefore, a comprehensive comparison of multiple models using Precision, Recall, F1-score, ROC-AUC, and PR-AUC is necessary to provide a more reliable assessment of fraud detection performance.

Another important gap concerns the limited comparison of sampling strategies. While the existing implementation applies under-sampling to address class imbalance, a systematic comparison with SMOTE is required to determine whether synthetic generation of minority-class samples can improve fraud detection. Furthermore, the presentation of experimental results requires improvement for scientific publication. Instead of numerous notebook screenshots showing data loading, dependency installation, statistical analysis, sampling, training, and evaluation, the methodology should be presented through concise workflow diagrams, structured tables, and publication-quality visualizations. Finally, the literature base should be more closely aligned with the research problem by emphasizing studies related to credit card fraud detection, imbalanced learning, SMOTE, ensemble machine learning, and anomaly detection. Addressing these gaps can improve the methodological rigor, clarity, and scientific contribution of the proposed fraud detection framework.

4. Proposed Methodology

The preprocessing stage consists of:

1. Dataset loading.
2. Dataset inspection.
3. Missing-value analysis.
4. Feature identification.
5. Target identification.
6. Statistical analysis.
7. Feature-target separation.
8. Train-test splitting.

The original implementation performs these operations before Logistic Regression training. The preprocessing stage is important because inconsistent or missing values can negatively affect model performance.

Exploratory Data Analysis

Exploratory analysis is performed to understand the distribution of the transaction data. The original study examines statistical measures and compares transaction characteristics. For the revised paper, the notebook screenshots should be replaced by:

- Class-distribution chart.
- Feature-distribution plots.
- Correlation matrix.
- Fraud-versus-legitimate comparison.
- Sampling comparison.

5. Handling Class Imbalance

5.1 Random Under-Sampling

Random under-sampling decreases the number of majority-class observations.

Advantages:

- Simple implementation.
- Reduced training time.
- Balanced training distribution.

Disadvantage:

- Potential loss of useful legitimate transaction information.

5.2 SMOTE

SMOTE generates synthetic minority-class samples. The synthetic observation can be expressed as:where is a minority sample, is one of its minority-class neighbours, and. Unlike under-sampling, SMOTE does not require discarding many majority observations.

6. Machine Learning Models

6.1 Logistic Regression

Logistic Regression is retained as the baseline model because it is simple, computationally efficient, and easy to interpret.

The probability of fraud is:

$$P(y=1|x) = \frac{1}{1+e^{-z}}$$

$$z = \beta_0 + \sum_{i=1}^{n} \beta_i x_i.$$

The original implementation trained Logistic Regression after preparing the training and testing data.

6.2 Random Forest

Random Forest consists of multiple decision trees.

For a classification problem, the final prediction can be represented as

$$\hat{y} = \mathrm{mode}(T_1(x), T_2(x), ..., T_n(x))$$

where represents an individual decision tree.

Random Forest is included because it can capture nonlinear relationships and interactions between transaction features

6.3 XGBoost

XGBoost is a gradient boosting algorithm.

Its objective function can be represented as:

$$Obj = \sum_i L(y_i, \hat{y}_i) + \sum_k \Omega(f_k)$$

where represents prediction loss and controls model complexity.

XGBoost is included because fraud patterns can involve nonlinear relationships that may not be captured effectively by Logistic Regression.

6.4 Autoencoder

The Autoencoder consists of an encoder and decoder.

The encoder generates a latent representation:

Z = fθ(x)

and the decoder reconstructs the input:

x^=gϕ(z).

The reconstruction error is:

E(x)=||x−x^||2

7. Experimental Configuration

The revised study should compare the models using the following experimental configurations.

Table 2 Experimental Design

Experiment	Data Condition	Model
E1	Original	Logistic Regression
E2	Under-Sampling	Logistic Regression
E3	SMOTE	Logistic Regression
E4	Original	Random Forest
E5	Under-Sampling	Random Forest
E6	SMOTE	Random Forest
E7	Original	XGBoost
E8	Under-Sampling	XGBoost
E9	SMOTE	XGBoost
E10	Original	Autoencoder

All sampling operations should be applied only to the training data after train-test splitting to avoid data leakage

8. Conclusion

This research presents an automated machine learning framework for credit card fraud detection with specific attention to class imbalance. The original implementation uses preprocessing, statistical analysis, under-sampling, train-test splitting and Logistic Regression and reports an accuracy of 93%. Although this baseline demonstrates the applicability of machine learning, accuracy alone does not adequately represent fraud-detection performance. A fraudulent transaction incorrectly classified as legitimate may have a substantially greater impact than a correctly classified legitimate transaction. Therefore, the revised framework evaluates Precision, Recall, F1-score, ROC-AUC and PR-AUC. The proposed methodology further extends the original implementation through comparison of Logistic Regression, Random Forest, XGBoost and Autoencoder approaches and through investigation of both under-sampling and SMOTE. The framework provides a systematic methodology for selecting a fraud detection model based on minority-class performance rather than global accuracy alone. The final experimental comparison should identify the model that provides the most appropriate balance between fraud detection, false-positive control, and computational efficiency. The proposed system provides a foundation for future development of adaptive, explainable, and real-time credit card fraud detection systems

Reference

[1]. 1st Chavda Dharmendra, 2nd Harshwardhansinh Chauhan, 3rd Krushal Gohil, 4th Vinayak Pandya, TIJER || ISSN 2349-9249 || © August 2026 Volume 13, Issue 8 || www.tijer.org, Credit Card Fraud Detection automation system Using Machine Learning

[2]. Harshwardhansinh K. Chauhan, Dr. Sheshang Degadwala, "Project Base Prediction Using Machine Learning and Deep Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN: 2456-3307, Volume 9, Issue 2, pp.22-29, March-April2023. Available at doi: https://doi.org/10.32628/CSEIT2390150 Journal URL: https://ijsrcseit.com/CSEIT2390150(2)

[3]. Harshwardhansinh K. Chauhan, Dr. Sheshang Degadwala, "Sensitivity Analysis of Project using Machine Learning ", International Journal of Scientific Research in Science, Engineering, and Technology (IJSRSET), Online ISSN: 2394-4099, Print ISSN: 2395- 1990, Volume 10 Issue 2, pp. 197-202, March-April 2023. Available at doi: https://doi.org/10.32628/IJSRSET2310128 Journal URL: https://ijsrset.com/IJSRSET2310128(3)

[4]. Harshwardhansinh K. Chauhan, Miss Kamini R. Parmar, Dr. Sheshang Degadwala, " Financial Risk Analysis using Machine Learning", International Journal of Scientific Research in Science and Technology (IJSRST), Print ISSN: 2395-6011, Online ISSN: 2395-602X, Volume 10, Issue 2, pp.352-358, March-April-2023. Available at doi: https://doi.org/10.32628/IJSRST2310117 Journal URL: https://ijsrst.com/IJSRST2310117(4)

[5]. Çavdar, E. Z., & Bozanta, A. (2026). AI-based credit card fraud detection: a machine learning approach with model explainability on real-world data. Knowledge and Information Systems.

[6]. Gamal, N., Younis, E. M. G., & Makram, W. M. (2026). Enhancing credit card fraud detection with a hybrid approach using machine and deep learning. Scientific Reports.

[7]. Becerra-Suárez, F. L., Arones-Pérez, P. P., Zamora-Del-Aguila, F. F., & Forero, M. G. (2026). Conservative plausibility-filtered SMOTE for credit card fraud detection under extreme class imbalance. Frontiers in Artificial Intelligence.

[8]. Applying supervised machine learning algorithms and ensemble models to enhance credit card fraud detection. Frontiers in Artificial Intelligence (2026).

[9]. Comparative Analysis of Machine Learning Algorithms and Data Balancing Techniques for Credit Card Fraud Detection. Procedia Computer Science (2024).