

American Journal of Economics and Business Management

Vol. 9 Issue 3 | **pp.** 801-810 | **ISSN:** 2576-5973

Available online @ https://www.globalresearchnetwork.us/index.php/ajebm

Article

Financial Fraud Detection in Business Entities: A Comparative Analysis of Machine Learning Methods

Kambarov Bakhriddin Panji ugli

1. PhD student at Tashkent State University of Economics

* Correspondence: b.qambarov@tseu.uz

Abstract: The rapid development of digital financial services has led to a significant increase in the volume of electronic transactions conducted by business entities, while simultaneously making the effective detection and management of financial fraud a growing concern. This issue becomes particularly challenging when fraudulent transactions represent only a very small proportion of overall operations, making the appropriate evaluation of classification models essential. This study comparatively investigates Logistic Regression, Random Forest, and XGBoost machine learning methods for financial fraud detection using highly imbalanced transactional data. To examine the impact of class imbalance on model performance, experiments were conducted under two conditions: with and without the application of the Synthetic Minority Over-sampling Technique (SMOTE). Model performance and consistency were evaluated using 3, 5, and 10 fold Stratified K-Fold cross-validation based on Precision, Recall, F1-score, ROC-AUC, and PR-AUC metrics. The final test results demonstrate that the effectiveness of SMOTE is not uniform across all machine learning models. In particular, Random Forest without SMOTE achieved an F1-score of 0.8743 and a PR-AUC of 0.8734. The findings provide a practical basis for improving automated monitoring of financial transactions and developing more effective financial fraud detection systems for business entities.

Keywords: financial fraud, machine learning, financial technologies, business entities, class imbalance, SMOTE, Logistic Regression, Random Forest, XGBoost, transaction detection.

Citation: Panji ugli, K. B. Financial Fraud Detection in Business Entities: A Comparative Analysis of Machine Learning Methods. American Journal of Economics and Business Management 2026, 9(3), 801-810

Received: 13th Dec 2025
Revised: 28th Jan 2026
Accepted: 15th Feb 2026
Published: 28th Mar 2026

Copyright: © 2026 by the authors. Submitted for open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/)

1. Introduction

The rapid development of the digital economy has significantly increased the volume and speed of transactions conducted through e-commerce, internet banking, and mobile payment systems [1] . The expansion of financial activity and the growth of household incomes play an important role in improving living standards and economic well-being, making the stability and security of financial transactions an important socioeconomic concern[2]. While this process provides greater convenience, it also exposes the financial ecosystem to the risks of cybercrime and fraud. Global payment card fraud losses exceeded USD 33.83–34 billion in 2023 [3], online payment losses are projected to reach USD 91 billion in 2028 [4]. Fraudsters use various techniques, such as card data skimming, phishing, creating fake websites, and stealing personal information [5], in particular, the share of losses associated with Card-Not-Present (CNP) transactions, in which the physical card is not present, is increasing [6].

Traditional systems rely primarily on static rules and specific features [7]. The dynamic and nonlinear nature of fraud patterns limits the adaptability of such methods [8], this increases the rate of false alarms, leading to the unjustified blocking of legitimate transactions [9]. Therefore, machine learning (ML), deep learning (DL), and explainable artificial intelligence (XAI) methods have become important areas of scientific and

practical research. However, the fact that fraudulent transactions account for less than 0.2% of all transactions creates bias during model training and remains a major methodological challenge in evaluating model performance due to the extreme class imbalance [10].

The literature review identifies three main research directions. First, modeling approaches: R.Qi emphasizes that traditional scoring models are unable to capture nonlinear relationships; M. N. Alatawi demonstrates the high stability of ensemble models; and Suresh and Anusha demonstrate the effectiveness of LSTM and CNN architectures alongside Random Forest, Gradient Boosting, and SVM. Second, addressing class imbalance: S. Jiang, R. Dong, J. Wang, and M. Xia, proposed the UAAD-FDNet model based on GANs and an attention mechanism, Y. Wu, L. Wang, H. Li, and J. Liu, applied SMOTE and ADASYN methods along with a CCNN model, while C. Huot, S. Heng, T. K. Kim, and Y. Han, achieved high sensitivity using a quantum autoencoder (QAE) [11]. Third, evaluation metrics: U. Porwal and S. Mukund, argued that ROC-AUC can artificially overestimate performance on highly imbalanced datasets and recommended using Precision-Recall and F1-score metrics; N. Baisholan, J. E. Dietz, S. Gnatyuk, M. Turdalyuly, E. T. Matson, and K. Baisholanova, emphasized the need to ensure result transparency in Random Forest and XGBoost models through threshold tuning and SHAP tools.

Existing studies show that the practical effectiveness of ML/DL models depends on properly addressing class imbalance, selecting appropriate evaluation metrics, and ensuring the interpretability of the results. Therefore, the aim of this study is to analyze and evaluate fraud transaction detection models while taking these aspects into account.

2. Materials and Methods

2.1. Dataset Description and Structure

The study used the Credit Card Fraud Detection benchmark dataset (https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud), which contains transactions made by European cardholders over two days in September 2013. The dataset comprises 284,807 transactions in total, of which only 492 are labeled as fraudulent. Fraudulent transactions account for 0.172% of all records, which reflects the extreme class imbalance observed in practice [12]. To ensure customer confidentiality and data privacy, most features were transformed using Principal Component Analysis (PCA); the structure of the dataset is presented in Table 1.

Table 1. Structure of the Credit Card Fraud Detection dataset

Attribute	Description	Type
V_1–V_{28}	28 anonymized attributes obtained via PCA (original features are undisclosed for confidentiality reasons)	Continuous
Time	Seconds elapsed between each transaction and the first transaction	Continuous
Amount	Transaction amount	Continuous
Class	Target variable: 1 — fraud, 0 — legitimate transaction	Binary (0/1)

The class distribution and the training/test split used in the experiments are given in Table 2. The ratio of legitimate to fraudulent transactions is approximately 578:1; therefore, plain accuracy is insufficient for model evaluation, since a model that labels every transaction as legitimate would still achieve an accuracy above 99.8%.

Table 2. Class distribution and training/test split

Class	Number of transactions	Share (%)	Training set (80%)	Test set (20%)
Legitimate (Class 0)	284,315	99.828	227,452	56,863
Fraud (Class 1)	492	0.172	394	98
Total	**284,807**	**100.000**	**227,846**	**56,961**

2.2. Data Preprocessing and Imbalance Handling

To ensure the objectivity and reliability of model evaluation, three preprocessing steps were applied (Table 3). First, the Time and Amount attributes, the only continuous features not transformed by PCA, were standardized using StandardScaler. This brought all features to a comparable scale and prevented variables with larger numerical ranges from dominating model parameters and gradient updates. Second, the dataset was divided into training (80%) and test (20%) subsets using stratified sampling, which preserved the original proportion of fraudulent transactions (0.172%) in both partitions. Third, to reduce the bias of the learning process toward the majority class, the minority (fraud) class was oversampled using the Synthetic Minority Over-sampling Technique (SMOTE), which generates synthetic fraud examples by interpolating between neighboring minority-class instances [13].

Table 3. Data preprocessing pipeline

Step	Method	Applied to	Purpose
Feature scaling	StandardScaler	Time, Amount	Equalize the scales of continuous features
Data splitting	Stratified sampling (80% / 20%)	Entire dataset	Preserve the fraud ratio (0.172%) in both subsets
Imbalance handling	SMOTE	Training set only	Reduce majority-class bias and avoid data leakage

To prevent data leakage and overly optimistic results, SMOTE was applied exclusively to the training partition. The test set retained its natural class distribution, so that the measured performance reflects real banking conditions, where fraudulent transactions are extremely rare.

2.3. Model Training and Baseline Algorithms

To establish baseline performance indicators for predicting the fraud risk of transactions, three classical algorithms of increasing representational capacity were selected: Logistic Regression, Random Forest, and eXtreme Gradient Boosting (XGBoost). This selection reflects the methodological shift documented in the literature, in which linear scoring models are progressively replaced by ensemble learners capable of capturing complex nonlinear dependencies. All models were trained on the same SMOTE-balanced training partition and evaluated on the same test set preserved in its natural imbalanced state, which ensured a fair comparison of the results under conditions approximating a real banking environment. The main characteristics of the selected algorithms are summarized in Table 4.

Table 4. Summary of the baseline classification algorithms

Algorithm	Type	Core mechanism	Strengths	Limitations
Logistic Regression	Linear, parametric	Sigmoid of a weighted sum of features	Interpretable and computationally efficient	Cannot capture nonlinear interactions
Random Forest	Bagging ensemble	Majority vote of decorrelated decision trees	Captures nonlinearity; robust to overfitting	Lower interpretability; larger memory footprint
XGBoost	Gradient boosting ensemble	Sequential trees minimizing a regularized objective	High accuracy with built-in regularization	Sensitive to hyperparameter settings

Logistic Regression serves as a linear parametric benchmark. It estimates the posterior probability that a transaction is fraudulent by passing a weighted linear combination of its features through the sigmoid function:

$$h_\theta(x) = \sigma(\theta^T x) = \frac{1}{1+e^{-\theta^T x}} \quad (1)$$

where x is the feature vector of a transaction (V_1, …, V_{28}, *Amount*, *Time*), θ is the vector of model weights, and $\sigma(\cdot)$ maps the linear score into the interval (0, 1). The weights are estimated by minimizing the binary cross-entropy (log-loss) function:

$$J(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i log\ h_\theta(x_i) + (1-y_i) log\left(1-h_\theta(x_i)\right)\right] \quad (2)$$

where N is the number of training instances and $y_i \in \{0, 1\}$ is the true class label. Because the decision boundary of the model is linear in the feature space, it offers high interpretability and computational efficiency, but it cannot capture complex nonlinear interactions between features. This limitation defines its role as a baseline rather than a standalone fraud detector [14].

Random Forest is a bagging (Bootstrap Aggregating) ensemble of decision trees. Each tree is trained on a bootstrap sample of the training data, and at every split only a random subset of features is considered, which decorrelates the individual trees and reduces the variance of the ensemble. The final decision of M trained trees is determined by majority voting:

$$\hat{y} = mode(\{h_m(x)\}_{m=1}^{M}) \quad (3)$$

where $h_m(x)$ is the prediction of the m-th decision tree. By aggregating many weakly correlated learners, the algorithm effectively captures nonlinear dependencies among features and remains robust to overfitting [15].

eXtreme Gradient Boosting is a gradient boosting ensemble in which decision trees are trained sequentially, with each new tree correcting the residual errors of the preceding ones. At iteration t, the algorithm minimizes the following regularized objective function:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{N} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \quad (4)$$

where l is the loss function, $\hat{y}_i^{(t-1)}$ is the prediction of the ensemble after $t-1$ iterations, f_t is the tree added at iteration t, and $\Omega(f_t)$ is a regularization term that restricts model complexity:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2 \quad (5)$$

where T is the number of leaves in the tree, w_j is the weight of the j-th leaf, and γ and λ are regularization coefficients. This built-in penalization of complexity allows XGBoost to achieve high predictive accuracy on tabular data while limiting overfitting, and the algorithm has been evaluated for fraud detection under the original class imbalance [16].

Thus, the three algorithms represent complementary modeling paradigms: a linear classifier, a variance-reducing parallel ensemble, and a bias-reducing sequential

ensemble. Their comparison will make it possible to determine the relative performance of these approaches for detecting fraudulent transactions under extreme class imbalance.

2.4. Evaluation Metrics

Model performance was assessed under the extreme class imbalance of the Credit Card Fraud Detection dataset, in which fraudulent transactions account for only 0.172% of all observations. Under such conditions, overall accuracy gives a misleadingly optimistic assessment, since a classifier that labels every transaction as legitimate would still exceed 99.8% accuracy. Therefore, PR-AUC and Recall were selected as the primary evaluation metrics, while Precision, F1-score, and ROC-AUC were used as supporting metrics (Table 5). This choice is consistent with evidence that ROC-AUC may overstate model performance on highly imbalanced data because of the overwhelming number of true negatives, and that the focus on AUC-PR and Recall is more informative for fraud detection [17].

All metrics are derived from the confusion matrix, which consists of true positives (TP, correctly identified fraudulent transactions), true negatives (TN, correctly identified legitimate transactions), false positives (FP, legitimate transactions incorrectly classified as fraudulent), and false negatives (FN, fraudulent transactions incorrectly classified as legitimate).

Table 5. Evaluation metrics and their roles in the study

Metric	Role	Rationale
PR-AUC	Primary	Threshold-independent; focuses on the minority fraud class and excludes true negatives
Recall	Primary	Measures the share of detected fraud; minimizes false negatives (undetected fraud)
Precision	Supporting	Controls false alarms, i.e., legitimate transactions incorrectly blocked
F1-score	Supporting	Balances Precision and Recall at a given decision threshold
ROC-AUC	Supplementary	Describes overall class separation; may be inflated by the large number of true negatives

Recall (Sensitivity) is the proportion of actual fraudulent transactions that are correctly identified by the model:

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

A low Recall value indicates that a considerable share of fraud remains undetected, and each false negative corresponds to a direct financial loss. For this reason, Recall was treated as the primary operational criterion, reflecting the ability of the model to minimize false negatives [18].

PR-AUC summarizes the trade-off between Precision and Recall across all classification thresholds. Unlike ROC-based measures, it does not involve true negatives and therefore focuses directly on the minority fraud class. It was computed as the average precision:

$$PR - AUC = \sum_n (R_n - R_{n-1}) P_n \quad (7)$$

where P_n and R_n are the Precision and Recall values at the n-th threshold. Because PR-AUC does not depend on a particular decision threshold, it allows models to be compared independently of threshold selection and was used as the primary criterion for comparing the models [19].

Precision is the proportion of transactions predicted as fraudulent that are actually fraudulent, and therefore reflects the number of false alarms generated by the system [20].

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

The F1-score is the harmonic mean of Precision and Recall and provides a balanced measure when both false positive and false negative errors matter:

$$F1-score = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN} (9)$$

Finally, ROC-AUC was calculated to evaluate the overall ability of the model to distinguish fraudulent from legitimate transactions across thresholds. The ROC curve relates the True Positive Rate (equal to Recall) to the False Positive Rate:

$$FPR = \frac{FP}{FP + TN} \quad (10)$$

Since the large number of true negatives keeps FPR low even when many legitimate transactions are misclassified, ROC-AUC was not considered sufficient as a standalone measure and was reported only as a supplementary indicator [21].

The combined use of these metrics enables the models to be evaluated from complementary perspectives: Recall measures the ability to detect fraud, PR-AUC summarizes performance on the minority class independently of the threshold, Precision and F1-score reveal the cost of detection in terms of false alarms, and ROC-AUC describes overall class separation.

3. Results

3.1. Classification Performance

The Logistic Regression, Random Forest, and XGBoost models, trained on the SMOTE-balanced training partition, were evaluated on the test set preserved in its natural imbalanced state (56,961 transactions, of which 98 were fraudulent). Since accuracy cannot objectively reflect model performance under extreme class imbalance, the analysis focuses primarily on PR-AUC and Recall, supported by Precision and F1-score. The results are presented in Table 6, where the best value in each column is highlighted in bold.

Table 6. Performance of the machine learning models on the test set

Model	Accuracy	Precision	Recall	F1-score	PR-AUC
Logistic Regression	0.9745	0.0581	**0.9184**	0.1092	0.7240
Random Forest	**0.9995**	**0.8842**	0.8571	**0.8705**	0.8512
XGBoost	0.9994	0.8737	0.8469	0.8601	**0.8720**

Logistic Regression achieved the highest Recall (91.84%), detecting 90 of the 98 fraudulent transactions, but its Precision was only 5.81%: roughly 6 of every 100 alerts corresponded to actual fraud. This indicates that the linear model generates an excessive number of false positives, which would sharply increase the operational costs of a bank. Random Forest captured the nonlinear dependencies in the data effectively and obtained the highest F1-score (87.05%), with a well-balanced Precision (88.42%) and Recall (85.71%). XGBoost achieved the highest PR-AUC (0.8720), the primary threshold-independent criterion, which indicates the most stable performance across decision thresholds, although its Recall (84.69%) and F1-score (86.01%) were slightly lower than those of Random Forest. Given that the test set contains only 98 fraudulent transactions, the differences between the two ensemble models are small and should be interpreted with caution.

3.2. Cost-Based Analysis

To assess the practical implications of the results, the confusion matrices were converted into expected financial loss, assuming a cost of 100 monetary units for each false negative (C_{FN}) and 2 monetary units for each false positive (C_{FP}). The total loss was calculated as:

$$TLoss = C_{FN} \cdot FN + C_{FP} \cdot FP \quad (11)$$

The results are given in Table 7.

Table 7. Confusion matrix errors and financial loss of the models (in monetary units)

Model	FN	FP	FN cost	FP cost	Total loss
Logistic Regression	8	1,452	800	2,904	3,704
Random Forest	14	11	1,400	22	**1,422**
XGBoost	15	12	1,500	24	1,524

Although Logistic Regression missed the fewest fraudulent transactions (8 FN), its 1,452 false alarms generated 2,904 monetary units in unnecessary operational costs and raised the total loss to 3,704. In contrast, Random Forest and XGBoost reduced the number of false positives to 11 and 12, respectively, lowering the total loss to 1,422 and 1,524 monetary units. Relative to Logistic Regression, this corresponds to cost reductions of 61.6% for Random Forest and 58.9% for XGBoost. For both ensemble models, the loss is dominated by the cost of missed fraud (FN), which indicates that further improvement should primarily target Recall.

Overall, the ensemble models substantially outperformed the linear baseline in Precision, F1-score, and total financial loss. XGBoost provided the best PR-AUC, while Random Forest showed the lowest total loss at the applied decision threshold.

4. Discussion

The findings demonstrate that the effectiveness of machine learning models for financial fraud detection cannot be assessed solely through conventional accuracy measures, particularly when the dataset is characterized by extreme class imbalance. Although all three models produced relatively high accuracy values, their ability to identify fraudulent transactions and control false alarms differed substantially. Logistic Regression achieved the highest Recall of 0.9184, indicating its strong ability to identify fraudulent transactions; however, its very low Precision of 0.0581 and F1-score of 0.1092 show that this advantage was accompanied by a large number of false-positive predictions. In practical financial monitoring systems, such a high false-positive rate may result in unnecessary transaction blocking, additional verification procedures, increased operational workload, and inconvenience for legitimate customers. Therefore, maximizing fraud detection sensitivity without considering Precision and the associated operational costs may not produce an economically efficient fraud detection system.

The substantially stronger performance of Random Forest and XGBoost confirms the importance of nonlinear ensemble learning for detecting complex fraud patterns. Random Forest achieved the highest Precision (0.8842) and F1-score (0.8705), while maintaining a Recall of 0.8571. This result suggests that the bagging mechanism and aggregation of multiple decision trees enable the model to capture complex interactions among transaction characteristics while controlling false-positive predictions. XGBoost, meanwhile, produced the highest PR-AUC of 0.8720, indicating stronger discrimination of fraudulent transactions across different classification thresholds. The advantage of ensemble approaches is consistent with previous studies showing that nonlinear and tree-based models can provide more effective fraud classification than conventional linear models when complex relationships exist among transaction features [7], [14], [15]. Nevertheless, the relatively small differences between Random Forest and XGBoost should be interpreted cautiously because the test set contained only 98 fraudulent transactions.

The results also highlight the importance of selecting evaluation metrics that correspond to the characteristics of fraud detection problems. Because fraudulent transactions represented only 0.172% of the dataset, accuracy remained between 0.9745 and 0.9995 despite substantial differences in Precision, Recall, F1-score, and PR-AUC. Consequently, accuracy alone could create an overly optimistic interpretation of model effectiveness. This observation supports previous arguments that ROC-AUC and

conventional accuracy measures may provide insufficient information when the minority class is extremely rare, whereas Precision-Recall measures and F1-score provide a more meaningful assessment of minority-class detection [9], [17]. In this context, PR-AUC is particularly useful because it evaluates the trade-off between Precision and Recall without being dominated by the large number of true-negative observations. Therefore, model selection for financial fraud detection should be based on multiple complementary metrics rather than a single overall performance indicator.

The cost-based analysis further demonstrates that statistical performance should be translated into operational and financial consequences. Logistic Regression generated only eight false negatives but produced 1,452 false positives, resulting in a total estimated loss of 3,704 monetary units. Random Forest and XGBoost reduced this loss to 1,422 and 1,524 monetary units, respectively, corresponding to reductions of 61.6% and 58.9% relative to Logistic Regression. These results indicate that a model with the highest Recall is not necessarily the most economically advantageous model. Random Forest provided the lowest total loss because its relatively high Recall was accompanied by substantially fewer false-positive classifications. This finding emphasizes that fraud detection represents a cost-sensitive classification problem in which false negatives and false positives have different economic consequences. Therefore, model deployment decisions should incorporate institution-specific fraud losses, investigation costs, transaction values, and customer-service consequences rather than relying exclusively on predictive metrics.

Another important implication concerns the use of SMOTE for handling class imbalance. Applying SMOTE only to the training partition while preserving the original class distribution in the test set provides a more realistic assessment of model behavior under actual operating conditions. This procedure also reduces the risk of data leakage and ensures that synthetic minority observations do not artificially improve test performance. Previous research has similarly emphasized that sampling and imbalance-handling strategies can substantially affect fraud detection performance [13], [14]. However, the present results also indicate that oversampling alone does not eliminate the trade-off between fraud detection and false alarms. Future model development should therefore combine resampling approaches with threshold optimization, cost-sensitive learning, alternative sampling strategies, and model calibration to determine whether further improvements in Recall can be achieved without substantially reducing Precision.

From a practical perspective, the findings suggest that Random Forest represents the most balanced option under the cost assumptions applied in this study, whereas XGBoost may be preferable when threshold-independent minority-class discrimination is prioritized. Rather than identifying a universally superior algorithm, the results demonstrate that the preferred model depends on the operational objective of the financial institution. Institutions seeking to minimize total detection costs may favor Random Forest, while those requiring greater flexibility in selecting decision thresholds may benefit from XGBoost because of its higher PR-AUC. Overall, the study confirms that effective automated fraud monitoring requires the joint consideration of class imbalance, model architecture, evaluation metrics, decision thresholds, and the asymmetric economic consequences of classification errors. Further validation using repeated cross-validation, temporal datasets, transaction-specific cost matrices, and explainability methods such as SHAP would strengthen the generalizability and practical applicability of these findings.

5. Conclusion

This study compared three classical machine learning models, namely Logistic Regression, Random Forest, and XGBoost, for credit card fraud detection under extreme class imbalance (0.172% fraudulent transactions). The models were trained on a SMOTE-balanced training partition and evaluated on a test set preserved in its natural imbalanced state, with PR-AUC and Recall as the primary evaluation metrics. The results lead to the following conclusions.

First, the ensemble models substantially outperformed the linear baseline. Random Forest and XGBoost reached Precision of 0.8842 and 0.8737, F1-score of 0.8705 and 0.8601, and PR-AUC of 0.8512 and 0.8720, respectively, whereas Logistic Regression obtained a Precision of only 0.0581, an F1-score of 0.1092, and a PR-AUC of 0.7240. Although Logistic Regression achieved the highest Recall (0.9184), it did so by generating 1,452 false alarms, which makes it impractical as a standalone detector. Second, no single ensemble model was superior on every criterion: XGBoost achieved the highest PR-AUC, while Random Forest showed the highest Precision and F1-score and the lowest total loss (1,422 versus 1,524 monetary units). Since the test set contains only 98 fraudulent transactions, the differences between these two models are small and should be interpreted with caution. Third, the cost-based analysis confirmed the practical value of the ensemble models, which reduced the total loss by 61.6% (Random Forest) and 58.9% (XGBoost) relative to Logistic Regression. For both ensembles, the remaining loss was dominated by missed fraud, which indicates that improving Recall is the main direction for further development. Finally, accuracy values between 0.9745 and 0.9995 did not reflect these substantial differences between the models, which confirms that PR-AUC, Recall, and cost-based measures should be used for evaluation instead of accuracy under extreme class imbalance.

Based on these findings, it is recommended that fraud detection models be evaluated primarily with PR-AUC, Recall, and cost-based loss measures rather than accuracy, that the decision threshold be optimized and SMOTE be combined with alternative imbalance-handling and cost-sensitive approaches in order to reduce missed fraud, that the cost matrix be defined from actual transaction amounts, that the results be confirmed through repeated cross-validation, temporal validation, and additional datasets, and that future studies be extended with explainability tools such as SHAP and with deep learning and anomaly detection.

REFERENCES

[1] S. E. Suresh and M. C. A. Y. Anusha, “International Journal of Research Publication and Reviews Credit Card Fraud Detection Using State of the Art Machine Learning and Deep Learning Algorithms,” vol. 6, no. 5, pp. 8669–8674, 2025.

[2] Tursunov Farxod Baxodir o'g'li, & Qambarov Bahriddin Panji o'g'li. (2024). Aholi turmush darajasi baholashda daromadlarning ahamiyati. Yashil Iqtisodiyot Va Taraqqiyot, 2(11), 11–15. https://doi.org/10.5281/zenodo.14241823.

[3] Y. Wu, L. Wang, H. Li, and J. Liu, “A Deep Learning Method of Credit Card Fraud Detection Based on Continuous-Coupled Neural Networks,” *Mathematics*, vol. 13, no. 5, pp. 1–18, 2025, doi: 10.3390/math13050819.

[4] N. Baisholan, J. E. Dietz, S. Gnatyuk, M. Turdalyuly, E. T. Matson, and K. Baisholanova, “FraudX AI: An Interpretable Machine Learning Framework for Credit Card Fraud Detection on Imbalanced Datasets,” *Computers*, vol. 14, no. 4, 2025, doi: 10.3390/computers14040120.

[5] L. Moumeni, M. Saber, I. Slimani, I. Elfarissi, and Z. Bougroun, “Machine Learning for Credit Card Fraud Detection,” *Lecture Notes in Electrical Engineering*, vol. 745, no. 24, pp. 211–221, 2022, doi: 10.1007/978-981-33-6893-4_20.

[6] G. K. Arun and P. Rajesh, “Analysis and prediction for credit card fraud detection dataset using data mining approaches,” *Int. J. Health Sci. (Qassim).*, vol. 6, no. April, pp. 4155–4173, 2022, doi: 10.53730/ijhs.v6ns5.9534.

[7] M. N. Alatawi, “Detection of fraud in IoT based credit card collected dataset using machine learning,” *Machine Learning with Applications*, vol. 19, no. December 2024, p. 100603, 2025, doi: 10.1016/j.mlwa.2024.100603.

[8] R. Qi, “An Empirical Study on Credit Risk Assessment Using Machine Learning: Evidence from the Kaggle Credit Card Fraud Detec-tion Dataset,” *Journal of Computer, Signal, and System Research*, vol. 2, no. 5, p. 48, 2025, [Online]. Available: https://doi.org/10.71222/7mn4wp34

[9] U. Porwal and S. Mukund, “Credit Card Fraud Detection in e-Commerce: An Outlier Detection Approach,” 2019, [Online]. Available: http://arxiv.org/abs/1811.02196

[10] S. Jiang, R. Dong, J. Wang, and M. Xia, “Credit Card Fraud Detection Based on Unsupervised Attentional Anomaly Detection Network,” *Systems*, vol. 11, no. 6, pp. 1–14, 2023, doi: 10.3390/systems11060305.

[11] C. Huot, S. Heng, T. K. Kim, and Y. Han, "Quantum Autoencoder for Enhanced Fraud Detection in Imbalanced Credit Card Dataset," *IEEE Access*, vol. 12, pp. 169671–169682, 2024, doi: 10.1109/ACCESS.2024.3496901.

[12] D. T. Mosa, S. E. Sorour, and A. A. Abohany, "CCFD : Efficient Credit Card Fraud Detection Using Meta-Heuristic Techniques and Machine Learning Algorithms," pp. 1–27, 2024.

[13] Z. Salekshahrezaee, J. L. Leevy, and T. M. Khoshgoftaar, "The effect of feature extraction and data sampling on credit card fraud detection," *J. Big Data*, vol. 10, no. 1, 2023, doi: 10.1186/s40537-023-00684-w.

[14] M. G. Wijaya, M. F. Pinaringgi, A. Y. Zakiyyah, and Meiliana, "Comparative Analysis of Machine Learning Algorithms and Data Balancing Techniques for Credit Card Fraud Detection," *Procedia Comput. Sci.*, vol. 245, pp. 677–688, 2024, doi: 10.1016/j.procs.2024.10.294.

[15] Y. Kumar, S. Saini, and R. Payal, "Comparative Analysis for Fraud Detection Using Logistic Regression, Random Forest and Support Vector Machine," *SSRN Electronic Journal*, vol. 7, no. 4, pp. 726–731, 2021.

[16] H. C. Du, L. Lv, H. Wang, and A. Guo, "A novel method for detecting credit card fraud problems," *PLoS One*, vol. 19, no. 3 March, pp. 1–26, 2024, doi: 10.1371/journal.pone.0294537.

[17] J. L. Leevy, J. M. Johnson, J. Hancock, and T. M. Khoshgoftaar, "Threshold optimization and random undersampling for imbalanced credit card data," *J. Big Data*, vol. 10, no. 1, 2023, doi: 10.1186/s40537-023-00738-z.

[18] N. A. Khasanovich, Z. F. Makhsetovich, J. N. Sobirovich, and M. A. Karamatdinovich, "Comparative efficiency analysis of cnn and vit models in brain cancer detection based on mri images," pp. 357–363, 2026.

[19] N. Lim and J. Lederer, "Efficient Feature Selection With Large and High-dimensional Data," *Proceedings of the 23rd International Conference on Machine learning -- ICML'06*, pp. 233–240, 2016, [Online]. Available: http://portal.acm.org/citation.cfm?doid=1143844.1143874%5Cnhttp://stats.stackexchange.com/questions/7207/roc-vs-precision-and-recall-curves%5Cnhttp://arxiv.org/abs/1609.07195

[20] N. A. Xasanovich, Z. F. Maxsetovic, and I. Kirish, "U- NET CHUQUR O ' QITIS H MODELI YORDAMIDA UMURTQA POG ' ONASINING KOMPYU TER TOMOGRAFIYASI TASVIRLARINI," vol. 4, no. 2, pp. 32–39, 2026.

[21] T. Albalawi and S. Dardouri, "Enhancing credit card fraud detection using traditional and deep learning models with class imbalance mitigation," *Front. Artif. Intell.*, vol. 8, no. October, pp. 1–10, 2025, doi: 10.3389/frai.2025.1643292.