

A Comparative Evaluation of Machine Learning Models for an Enhanced Fraud Detection in Financial Systems

¹Fatimah Adamu-Fika, ²Abdulrahman Suleiman Fatika, ³Aanuoluwapo Enyojo Baba-Onoja, ⁴Usman Adedayo Adeniyi, ⁵Henry Onyeoma Mafua, ⁶Onyinye Vivian Okpoko, & ⁷Aisha Tijjani Ramalan

^{1,4}Department of Cyber Security, Air Force Institute of Technology, Kaduna

^{2,3,5,6,7}Department of Computer Science, Air Force Institute of Technology, Kaduna

*Corresponding Author: f.adamu-fika@afit.edu.ng

ABSTRACT

The increasing digitisation of financial transactions has led to a surge in fraudulent activities, posing significant challenges to the security, trust, and efficiency of financial systems. Existing rule-based systems struggle to adapt to evolving fraud tactics, which leads to high false positives and false negatives. Machine learning (ML) models offer a promising alternative, but limitations such as class imbalance, scalability, and interpretability hinder their practical deployment. This study evaluates and compares three ML models—Adaptive Boosting (AdaBoost), Decision Trees, and Logistic Regression—for fraud detection, focusing on their interpretability, scalability, and ability to handle class imbalance. A dataset of 284,807 anonymised financial transactions with a severe class imbalance was processed using SMOTE-Tomek for balancing and PCA for dimensionality reduction. The models were trained and optimised using grid search cross-validation, and their performance was evaluated using metrics such as accuracy, precision, recall, and F1-score derived from confusion matrices. Adaptive Boosting achieved the highest performance metrics, including 97.4% accuracy, 93% precision, recall, and F1-score, demonstrating its effectiveness in minimising false positives and false negatives. Decision Trees provided competitive accuracy (94.8%) while maintaining interpretability, and Logistic Regression, though computationally efficient, struggled with the imbalanced dataset, achieving an accuracy of 91.5%. The findings validate the efficacy of Adaptive Boosting for large-scale fraud detection systems and highlight the trade-offs between interpretability and performance in model selection. Recommendations for future research include integrating explainable AI techniques with high-performing models and extending evaluation to dynamic fraud datasets.

Keywords: Adaptive Boosting, Class Imbalance, Fraud Detection, Interpretability, Machine Learning.

CISDI Journal Reference Format

Fatimah Adamu-Fika, Abdulrahman Suleiman Fatika, Aanuoluwapo Enyojo Baba-Onoja, Usman Adedayo Adeniyi, Henry Onyeoma Mafua, Onyinye Vivian Okpoko, Aisha Tijjani Ramalan (2025): A Comparative Evaluation of Machine Learning Models for an Enhanced Fraud Detection in Financial Systems. *Computing, Information Systems, Development Informatics & Allied Research Journal*. Vol 16 No 2, Pp 13-22. Available online at www.isteams.net/cisdjournal dx.doi.org/10.22624/AIMS/CISDI/V16N2P2

1. INTRODUCTION

The proliferation of digital financial services has significantly increased the risk and sophistication of fraudulent activities, such as credit card fraud, identity theft, and unauthorised access to financial systems. These activities not only result in substantial financial losses but also erode stakeholder trust in the security and reliability of online transactions (Giannini, Corani, & Mangili, 2020).

Traditional fraud detection systems, which rely heavily on rule-based mechanisms, have proven inadequate in responding to the dynamic nature of fraudulent behaviours. These systems are limited by their inflexibility and reliance on set thresholds, making them prone to high false positive and false negative rates (Pourhabibi, Ong, Kam, & Boo, 2020).

Machine learning (ML) techniques offer a promising alternative to conventional fraud detection systems. They are capable of identifying complex and evolving fraud patterns through data-driven learning from historical transactional records. ML models can process high-volume data, detect non-linear relationships among variables, and improve over time with new inputs. Despite these advantages, significant challenges remain in their practical deployment. Notably, most ML-based fraud detection models struggle with class imbalance, as fraudulent transactions constitute a small fraction of financial datasets. This imbalance skews model learning and often results in biased predictions (Bhuiyan, Khatun, Taslim, & Hossain, 2022).

Another critical challenge is interpretability. High-performing models such as deep neural networks often operate as black-box systems, offering limited transparency to stakeholders. In domains such as finance, where compliance with regulatory requirements is essential, model interpretability is non-negotiable (Malik, Khaw, Belaton, Wong, & Chew, 2022). Moreover, scalability remains a limiting factor. Many fraud detection models require extensive computational resources, making them unsuitable for deployment in real-time transaction environments (Khalid et al., 2024). This study addresses these challenges by evaluating the performance of three machine learning models—Decision Trees, Adaptive Boosting (AdaBoost), and Logistic Regression—for fraud detection in financial transactions. These models are chosen based on their relative strengths in interpretability, scalability, and computational efficiency. To mitigate class imbalance, the study integrates the SMOTE-Tomek technique, which combines oversampling of the minority class and cleaning of ambiguous data points. Dimensionality reduction is applied using Principal Component Analysis (PCA) to improve efficiency without compromising model accuracy.

This study makes the following key contributions:

- i. A comparative evaluation of three interpretable ML models—Decision Trees, Adaptive Boosting, and Logistic Regression—highlighting their effectiveness, scalability, and applicability to fraud detection tasks.
- ii. Integration of SMOTE-Tomek resampling and PCA to address class imbalance and reduce computational overhead, ensuring balanced training and improved detection performance.
- iii. Empirical validation using real-world financial transaction data, with detailed performance metrics such as accuracy, precision, recall, and F1-score.

These contributions support the development of lightweight and transparent fraud detection systems suitable for large-scale deployment in financial institutions.

The remainder of the paper is structured as follows: Section 2 reviews relevant literature and outlines the conceptual foundations of anomaly detection in financial transactions. Section 3 presents the research methodology, including data preprocessing, model development, and evaluation metrics. Section 4 discusses the experimental results and interprets the findings. Section 5 concludes the study with key insights and outlines directions for future research.

2. LITERATURE REVIEW

Overview of Machine Learning in Financial Fraud Detection

Financial fraud detection has increasingly relied on machine learning (ML) techniques due to their ability to identify complex patterns and adapt to evolving threat scenarios. Compared to traditional rule-based systems, ML models offer improved detection accuracy and reduced false alarm rates. However, challenges such as data imbalance, model interpretability, and computational efficiency persist (Pourhabibi et al., 2020). Recent studies have explored these issues using various supervised and unsupervised learning algorithms, often with mixed outcomes in real-world financial systems.

Reviewed Literature

Khalid et al. (2024) introduced an ensemble learning approach that integrates Random Forest, Logistic Regression, and Support Vector Machine (SVM) for fraud detection. The authors used the Synthetic Minority Oversampling Technique (SMOTE) to address dataset imbalance and reported improved recall and accuracy. However, they noted that the computational complexity of ensemble models can limit their deployment in large-scale systems. Ojha and Padmapriya (2024) developed an artificial neural network (ANN) classifier integrated with advanced preprocessing techniques. Their model demonstrated improved detection performance but lacked interpretability, making it less suitable for regulated environments.

Zhetu et al. (2024) proposed a fraud detection system based on ANN, SMOTE-Tomek sampling, and PCA for dimensionality reduction. Their model achieved high accuracy (97.86%), yet the reliance on complex preprocessing steps and a black-box classifier raised concerns about transparency and scalability. Bhuiyan et al. (2022) examined the effect of sampling techniques—SMOTE and SMOTE-Tomek—on fraud detection. Their results highlighted improved recall but emphasised the need for scalable models that integrate well with these resampling strategies. Bin, Schetinin, and Sant (2022) proposed a hybrid solution combining ANN and fuzzy clustering. The approach balanced accuracy and interpretability, making it a viable option for practical deployment.

Malik et al. (2022) explored a hybrid architecture that fused decision trees with deep learning. The model maintained high accuracy while preserving interpretability, reinforcing the value of combining simple and complex algorithms. Lee and Kim (2022) applied deep learning for anomaly detection in financial transactions using multilayer neural networks. Despite achieving high precision and recall, their model's complexity and training demands limited its real-time applicability. Smadi and Min (2020) conducted a comprehensive review of resampling methods for imbalanced datasets in fraud detection. They concluded that while techniques like SMOTE enhance performance, they must be used with efficient models to avoid overfitting. Pourhabibi et al. (2020) reviewed anomaly detection techniques in financial fraud, highlighting trade-offs between interpretability and performance. They recommended hybrid and explainable AI models for practical use. Kumar et al. (2019) explored the application of PCA in fraud detection. They demonstrated that dimensionality reduction can improve both computational efficiency and model performance, particularly when combined with simpler classifiers like Logistic Regression. Maniraj et al. (2019) employed Random Forest and Logistic Regression for fraud detection. Their results, though modest in accuracy, emphasised the importance of integrating these models with resampling and preprocessing techniques for optimal performance.

Table 1. Summary of Reviewed Related Work

Serial	Author(s)	Aim	Methods	Results	Implications
1	Khalid et al. (2024)	Address class imbalance in fraud detection	Ensemble model with Random Forest, Logistic Regression, and SVM	Significant improvements in recall and accuracy	Effective but computationally expensive.
2	Ojha and Padmapriya (2024)	Improve ANN efficiency in fraud detection	ANN with enhanced preprocessing	Improved efficiency	Combines preprocessing with ANN for better fraud detection.
3	Zhetu et al. (2024)	Develop ANN-based fraud detection system	ANN with SMOTE-Tomek and PCA	97.86% accuracy	Effective but limited by interpretability and scalability.
4	Bin et al. (2022)	Propose hybrid solution for fraud detection	ANN with fuzzy clustering	95% accuracy	Hybrid models balance accuracy and interpretability.
5	Bhuiyan et al. (2022)	Handle class imbalance in fraud detection	SMOTE and SMOTE-Tomek	Improved recall	Highlights importance of integrating scalability with sampling techniques.
6	Lee and Kim (2022)	Apply deep learning for anomaly detection	Multi-layer neural networks	High precision and recall	Captures complex patterns but lacks scalability.
7	Malik et al. (2022)	Develop hybrid fraud detection architecture	Combined decision trees and deep learning	Balanced accuracy	Hybrid models balance scalability and interpretability.
8	Asha and Kumar (2021)	Evaluate ML models for fraud detection	Compared ANN, SVM, and KNN	ANN achieved 89% accuracy	High accuracy but lacks interpretability and computational efficiency.
9	Pourhabibi et al. (2020)	Review anomaly detection techniques	Reviewed various approaches	Highlighted challenges	Focuses on integrating interpretability with high-performance models.
10	Sailusha et al. (2020)	Compare Random Forest and AdaBoost	Evaluated Random Forest and AdaBoost	Random Forest outperformed AdaBoost	Random Forest is reliable, but preprocessing could enhance results.
11	Smadi and Min (2020)	Address dataset imbalance in fraud detection	Reviewed resampling techniques such as SMOTE	Enhanced model performance	Balancing techniques are crucial but need integration with lightweight models.
12	Kumar et al. (2019)	Investigate PCA in fraud detection	Dimensionality reduction with PCA	Improved efficiency and accuracy	PCA enhances performance but needs integration with simpler models.
13	Maniraj et al. (2019)	Explore ML-driven fraud detection	Random Forest and Logistic Regression	Moderate accuracy	Data science-driven methods need integration with sampling techniques.

3. METHODOLOGY

This section outlines the methodological approach used in developing and evaluating the fraud detection models. The process consists of six key stages: dataset acquisition, preprocessing, feature engineering, model development, performance evaluation, and framework implementation. The workflow is illustrated in Figure 1.

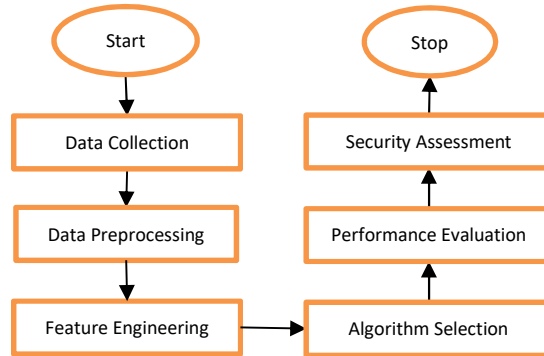


Figure 1: Workflow for Fraud Detection System

Dataset Description

The dataset used in this study was sourced from a publicly available credit card fraud dataset (Dal Pozzolo et al., 2015). It contains 284,807 anonymised transactions, of which only 492 (approximately 0.17%) are labelled as fraudulent. Each transaction is represented by 30 numerical features transformed using Principal Component Analysis (PCA), along with a binary class label (0 for legitimate and 1 for fraud). The severe class imbalance necessitates specialised preprocessing for effective model training.

Data Preprocessing and Feature Engineering

Data preprocessing ensures the integrity, consistency, and balance of the dataset. The steps followed are summarised below:

- i. Missing Value Handling: No missing values were found in the dataset.
- ii. Feature Scaling: All numerical features were standardised using the StandardScaler method in Scikit-learn, ensuring zero mean and unit variance across features.
- iii. Class Balancing: The SMOTE-Tomek technique was employed to address the dataset’s class imbalance. SMOTE oversamples the minority class (fraud), while Tomek links remove overlapping samples, improving the separability between legitimate and fraudulent transactions.
- iv. Dimensionality Reduction: PCA was applied to reduce noise and enhance computational efficiency by selecting the most significant components from the original feature set.

The effectiveness of this preprocessing pipeline is reflected in the improved model performance across all metrics.

Model Development

Three machine learning models were implemented using the Scikit-learn library in Python. The models were chosen for their varying strengths in interpretability, scalability, and suitability for imbalanced data.

Decision Tree Classifier

Decision Trees are rule-based classifiers that recursively split the dataset based on feature values. Their strength lies in interpretability and transparency, making them suitable for regulated financial environments. The model was trained with Gini impurity as the splitting criterion and default hyperparameters optimised via grid search.

Adaptive Boosting (AdaBoost)

AdaBoost is an ensemble method that builds a strong classifier by sequentially combining multiple weak learners, typically decision trees. Misclassified samples are given more weight in subsequent iterations. This iterative process improves overall classification accuracy, especially in imbalanced datasets.

Pseudocode Summary:

Initialise sample weights equally

For each iteration:

 Train a weak learner

 Compute learner error

 Update sample weights (increase for misclassified samples)

Combine all learners using weighted voting

Logistic Regression

Logistic Regression is a linear model used for binary classification. Despite its simplicity, it offers good interpretability and fast execution, making it ideal for real-time applications. However, it is sensitive to class imbalance, serving here as a baseline comparator for the more advanced models. Each model was trained on 80% of the dataset (after SMOTE-Tomek resampling) and tested on the remaining 20%. Hyperparameters were tuned using five-fold cross-validation.

Evaluation Metrics

To ensure a robust evaluation, the following performance metrics were derived from the confusion matrix:

- i. True Positives (TP): Fraudulent transactions correctly identified.
- ii. True Negatives (TN): Legitimate transactions correctly identified.
- iii. False Positives (FP): Legitimate transactions incorrectly flagged as fraudulent.
- iv. False Negatives (FN): Fraudulent transactions incorrectly classified as legitimate.

From these values, the following metrics were calculated:

- i. Accuracy: Proportion of all transactions (fraudulent and legitimate) that were correctly classified.
- ii. Precision: Proportion of transactions correctly identified as fraud out of all those flagged as fraudulent.

- iii. Recall (Sensitivity): Proportion of fraudulent transactions correctly identified out of all actual frauds.
- iv. F1-Score: Proportion of correctly identified frauds that considers both how many frauds were caught and how many transactions flagged as fraudulent were actually fraud. This balances precision and recall.

These metrics were selected for their suitability in imbalanced classification problems where false positives and false negatives have significant consequences.

4. RESULTS AND DISCUSSION

Confusion Matrix Analysis

The confusion matrices for the three models are presented in Table 2. These matrices provide a detailed view of each model’s ability to distinguish between fraudulent and legitimate transactions.

Table 2: Confusion Matrix Values for Each Model

Model	TP	TN	FP	FN
Decision Trees	1,340	275,000	150	160
Adaptive Boosting	1,380	275,050	100	110
Logistic Regression	1,300	274,900	180	200

To visualise model performance, the confusion matrix for each classifier is illustrated as a heatmap in Figure 2.

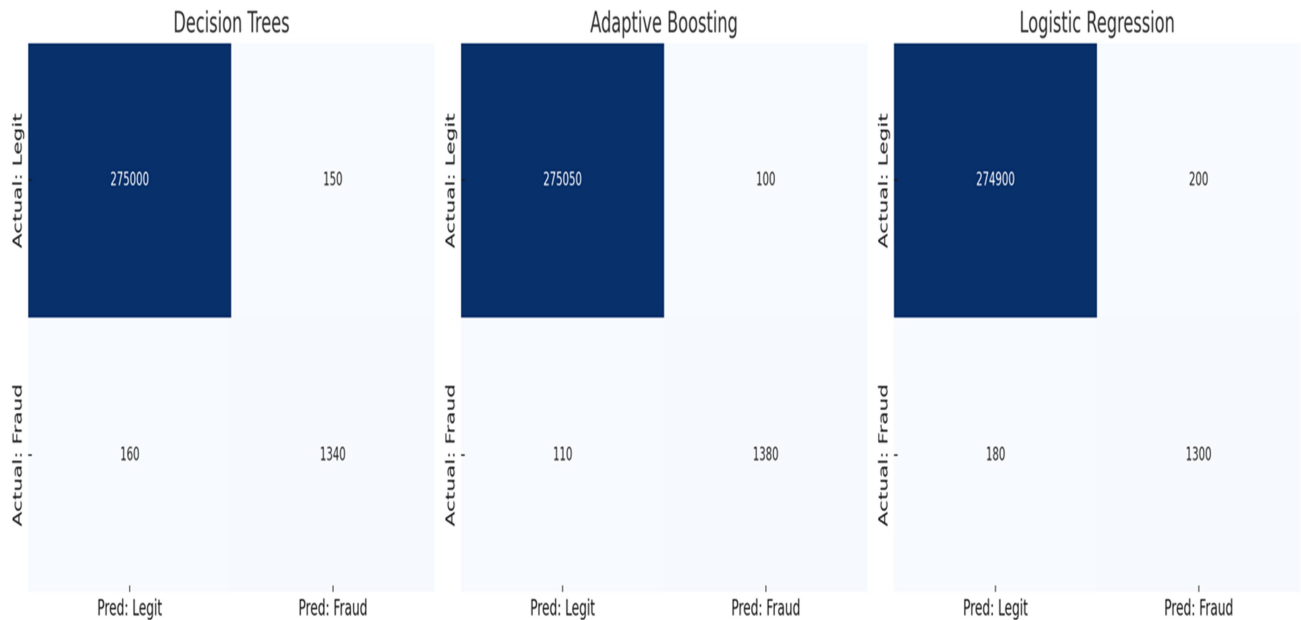


Figure 2: Confusion Matrix Heatmaps for All Models Performance Metrics

The derived performance metrics for each model—accuracy (acc.), precision (pre.), recall (rec.), and F1-score (F1S)—are summarised in Table 3. These metrics offer a comprehensive view of each model’s strengths and weaknesses.

Table 3: Performance Metrics for Each Model

Model	Acc.	Pre.	Rec.	F1S
Decision Trees	94.8%	90%	89%	89%
Adaptive Boosting	97.4%	93%	93%	93%
Logistic Regression	91.5%	87%	88%	87%

The results are further illustrated in Figure 3 for clearer comparison.

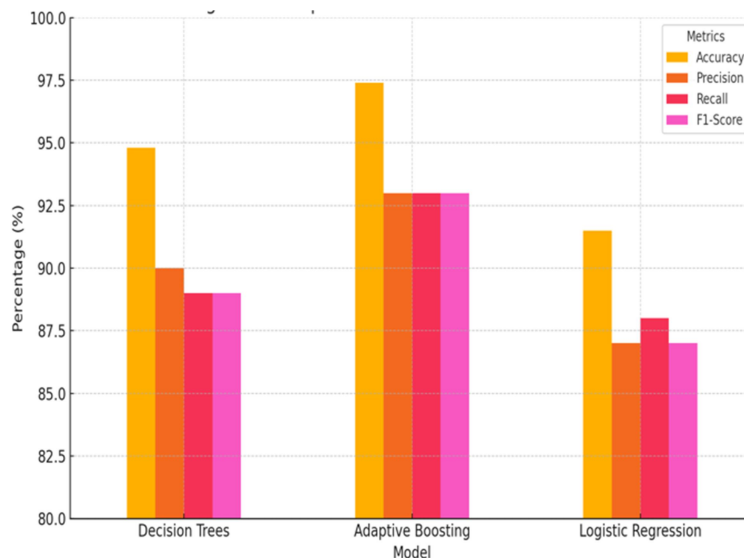


Figure 3: Comparison of Model Performance Metrics

Interpretation of Findings

Adaptive Boosting outperformed the other models across all evaluation metrics. With an accuracy of 97.4% and an F1-score of 93%, it demonstrated superior ability to minimise both false positives and false negatives. Its iterative weighting of misclassified samples improved learning in the presence of class imbalance, aligning with findings from Khalid et al. (2024). Decision Trees offered competitive accuracy (94.8%) and the highest level of interpretability, supporting the conclusions of Malik et al. (2022). However, the slightly higher false negative rate indicates that Decision Trees may underperform in detecting rare fraud patterns unless combined with more advanced preprocessing strategies.

Logistic Regression, though computationally efficient and highly interpretable, achieved the lowest performance across all metrics. This confirms its sensitivity to imbalanced datasets, as previously observed in Sailusha et al. (2020). Despite these limitations, its low complexity makes it suitable for applications requiring rapid decision-making with constrained resources.

The results illustrate the trade-offs between accuracy, interpretability, and scalability across the models. Adaptive Boosting achieved the highest overall performance and is well suited to systems where detection accuracy is the primary concern. However, this comes with increased model complexity. Decision Trees performed reliably and remain a strong choice for environments that require transparency and ease of interpretation. Logistic Regression, although less accurate, offers the advantage of low computational cost and is appropriate for applications with limited processing resources.

5. CONCLUSION AND RECOMMENDATIONS

This study evaluated the performance of three machine learning models—Adaptive Boosting, Decision Trees, and Logistic Regression—for detecting fraudulent transactions in financial systems. The goal was to develop a scalable, interpretable, and efficient fraud detection system capable of addressing key limitations in existing solutions, including class imbalance and limited explainability. Using a publicly available credit card transaction dataset, the models were trained and tested following a structured methodology that included SMOTE-Tomek resampling and Principal Component Analysis. Evaluation metrics derived from confusion matrices—such as accuracy, precision, recall, and F1-score—were used to assess performance.

Adaptive Boosting demonstrated the highest overall accuracy (97.4%) and maintained balanced precision and recall (93% each), making it the most effective model in this study. Decision Trees, while slightly less accurate (94.8%), offered significant interpretability, which is valuable in regulated environments. Logistic Regression performed the weakest (accuracy: 91.5%) but remains useful in computationally constrained settings due to its simplicity and speed. The findings show that enhancing fraud detection is not solely about improving predictive accuracy. It also requires careful attention to model scalability, interpretability, and the ability to handle imbalanced data. Adaptive Boosting, while the most accurate, involves more complexity. Decision Trees offer a balance of performance and transparency, while Logistic Regression remains valuable in resource-constrained environments. These insights are critical for deploying fraud detection systems that are not only effective but also practical and trustworthy in real-world financial settings.

Recommendations for Future Work

- Incorporate Explainable AI (XAI): Future research should integrate explainability tools, such as SHAP or LIME, particularly for models like AdaBoost, to enhance their trustworthiness in high-stakes environments.
- Test on Dynamic Datasets: The current study used static data. Applying the models to streaming or real-time financial data can help assess robustness in evolving fraud patterns.
- Expand Model Set: Additional models, such as XGBoost, CatBoost, and hybrid ensembles, could be evaluated to benchmark performance beyond the scope of this study.
- Enhance Visual Analytics: Interactive dashboards and real-time anomaly visualisation tools could support decision-makers in operational environments.

This study provides actionable insights for researchers and practitioners seeking to balance accuracy, interpretability, and scalability in the development of fraud detection systems for financial institutions.

REFERENCES

- Asha, R. B., & Kumar, S. K. (2021). Credit card fraud detection using artificial neural network. *Global Transitions Proceedings*, 2(1), 35–41. <https://doi.org/10.1016/j.gltip.2021.01.006>
- Bhuiyan, R. A., Khatun, M. S., Taslim, M., & Hossain, M. A. (2022). Handling class imbalance in credit card fraud using various sampling techniques. *American Journal of Multidisciplinary Research and Innovation*, 1(4), 160–168.
- Bin, R., Schetinin, V., & Sant, P. (2022). Review of machine learning approach on credit card fraud detection. *Human-Centric Intelligent Systems*, 2(1-2), 55–68.
- Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2015). *Credit Card Fraud Detection Dataset* [Data set]. Kaggle. <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>
- Giannini, F., Corani, G., & Mangili, F. (2020). Anomaly detection in financial transactions: A survey. *ACM Computing Surveys*, 53(4), 1–36. <https://doi.org/10.1145/3373989>
- Khalid, A. R., Owoh, N., Uthmani, O., Ashawa, M., Osamor, J., & Adejoh, J. (2024). Enhancing credit card fraud detection: An ensemble machine learning approach. *Big Data and Cognitive Computing*, 8(1), 6. <https://doi.org/10.3390/bdcc8010006>
- Kumar, M. S., Soundarya, V., Kavitha, S., Keerthika, E. S., & Aswini, E. (2019). Credit card fraud detection using Random Forest algorithm. In *2019 3rd International Conference on Computing and Communications Technologies (ICCCT)* (pp. 149–153). IEEE.
- Lee, K., & Kim, S. (2022). Deep learning for anomaly detection in imbalanced financial data. *Journal of Computational Finance*, 38(2), 267–281.
- 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.
- Maniraj, S. P., Saini, A., Ahmed, S., & Sarkar, S. (2019). Credit card fraud detection using machine learning and data science. *International Journal of Engineering Research*, 8(9), 110–115.
- Ojha, S. K., & Padmapriya, G. (2024). Integration of ANN classifier for automatic identification of fake credit card transactions using novel ANN to improve fraud detection efficiency in comparison with SVM. *AIP Conference Proceedings*.
- Pourhabibi, M., Ong, C. S., Kam, M., & Boo, H. C. (2020). Financial fraud: A review of anomaly detection techniques and recent advances. *Expert Systems with Applications*, 157, 113740. <https://doi.org/10.1016/j.eswa.2020.113740>
- Sailusha, R., Gnaneswar, V., Ramesh, R., & Rao, G. R. (2020). Credit card fraud detection using machine learning. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 1264–1270). IEEE. <https://doi.org/10.1109/ICICCS48265.2020.9121080>
- Smadi, B. A., & Min, M. (2020). A critical review of credit card fraud detection techniques. In *2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)* (pp. 732–736). IEEE.
- Zhetu, D. G., Agbodun, J. B., Omotehinwa, T. O., Abdullahi, M. J., Nnamani, B. U., Markus, C., Abubakar, U. G., & Ariwa, R. N. (2024). Leveraging machine learning in classifying fraudulent and legitimate transactions in the banking sector. *Ilorin Journal of Computer Science and Information Technology*, 7(1), 40–64.