iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** I **Month of publication:** January 2026

DOI: https://doi.org/10.22214/ijraset.2026.76879

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue I Jan 2026- Available at www.ijraset.com

A Study on Financial Fraud Detection using Explainable AI

Happy Mistry
III Semester M. tech, Gandhinagar Institute of Technology, Gandhinagar

Abstract: ***With the rapid growth of digital banking, online payments, and cashless transactions, financial fraud has become a major challenge for banks and financial institutions. Fraudsters continuously develop new techniques, making traditional rule-based fraud detection systems ineffective and outdated. In recent years, machine learning and deep learning models have shown strong potential in identifying fraudulent transactions by learning complex patterns from large volumes of financial data. However, most of these models function as black-box systems, meaning their decisions are difficult to understand and explain. This lack of transparency creates trust issues for users and raises concerns regarding regulatory compliance in the financial sector. This study focuses on financial fraud detection using a combination of stacking ensemble learning and Explainable Artificial Intelligence (XAI) techniques. By integrating multiple high-performing machine learning models, the stacking approach improves detection accuracy and handles challenges such as class imbalance and evolving fraud patterns. At the same time, explainability methods such as SHAP, LIME, and feature importance analysis are used to provide clear insights into how and why a transaction is classified as fraudulent or legitimate. The study highlights that combining high accuracy with interpretability is both achievable and necessary for modern fraud detection systems. By improving transparency and trust, explainable fraud detection models can better meet real-world operational and regulatory requirements.***

Keywords: ***Financial Fraud Detection, Explainable Artificial Intelligence (XAI), Stacking Ensemble Learning, Machine Learning, Credit Card Fraud, Class Imbalance, SHAP, LIME, Model Interpretability, Regulatory Compliance, Transaction Analysis, Ethical AI.***

I. INTRODUCTION

The rapid advancement of digital technologies has transformed the financial sector by enabling fast, convenient, and cashless transactions. Online banking, mobile payments, credit cards, and digital wallets have become an essential part of everyday life. While these technologies have improved efficiency and customer experience, they have also increased the risk of financial fraud. Fraudsters continuously exploit system vulnerabilities, leading to significant financial losses, damage to institutional reputation, and loss of customer trust. As the volume and complexity of digital transactions continue to grow, detecting fraudulent activities in an accurate and timely manner has become a critical challenge for financial institutions.

Traditional fraud detection systems mainly rely on predefined rules and manual checks. Although such systems are simple to implement, they are limited in their ability to detect new and evolving fraud patterns. Fraud techniques change rapidly, making static rule-based approaches ineffective and prone to high false positive rates. To overcome these limitations, machine learning techniques have been widely adopted for fraud detection. These models can analyze large volumes of transaction data, learn hidden patterns, and adapt to new fraud behaviors over time. More recently, deep learning and ensemble learning techniques have further improved detection accuracy by capturing complex and non-linear relationships within data.

Despite their strong performance, most advanced machine learning and deep learning models operate as black-box systems. They provide predictions without clearly explaining the reasoning behind their decisions. In the financial domain, where automated decisions can directly impact customers, this lack of transparency creates serious concerns. Financial institutions are often required to justify why a transaction was flagged as fraudulent, especially under strict regulatory frameworks and compliance requirements. Without clear explanations, even highly accurate models may fail to gain trust from analysts, customers, and regulatory authorities.

Explainable Artificial Intelligence (XAI) has emerged as a solution to address this challenge. XAI techniques aim to make model predictions understandable by explaining how different features influence the final decision. By integrating explainability with high-performance models, it becomes possible to build fraud detection systems that are both accurate and transparent. This work focuses on the use of stacking ensemble learning combined with Explainable AI techniques to achieve reliable fraud detection while maintaining interpretability. Such an approach not only improves detection performance but also supports trust, accountability, and ethical deployment of AI in modern financial systems.

©IJRASET: All Rights are Reserved

II. LITERATURE REVIEW

This section presents a comprehensive review of existing research related to financial fraud detection, machine learning-based approaches, ensemble models, and Explainable Artificial Intelligence (XAI). The reviewed studies focus on improving fraud detection accuracy while addressing challenges such as transparency, trust, class imbalance, and regulatory compliance.

A. Hasan et al. (2024) – Explainable AI in Credit Card Fraud Detection

Hasan et al. (2024) proposed an Explainable Artificial Intelligence–based framework for detecting credit card fraud with a strong focus on transparency and interpretability. The study highlighted that financial institutions increasingly depend on automated decision-making systems, yet the lack of explanation behind fraud predictions often creates trust issues among analysts and customers. The authors employed traditional machine learning models and enhanced them using SHAP-based explanations to understand feature contributions at both global and local levels. The dataset used in the study consisted of real-world credit card transactions with a highly imbalanced class distribution, which is a common challenge in fraud detection tasks.

The results demonstrated that explainability significantly improved analysts' understanding of fraud patterns and reduced blind reliance on model predictions. However, the study primarily focused on offline evaluation and did not explore real-time transaction processing. Additionally, while accuracy improvements were reported, the paper did not extensively compare ensemble models with deep learning techniques. This work is valuable for emphasizing the importance of explainability, but it leaves scope for further research on scalability and real-world deployment.

B. Almalki and Masud (2025) – Financial Fraud Detection using XAI & Stacking Ensemble

Almalki and Masud (2025) introduced a robust stacking ensemble model for financial fraud detection by combining XGBoost, LightGBM, and CatBoost classifiers. The study addressed two major challenges in fraud detection: achieving high prediction accuracy and ensuring model interpretability. The IEEE-CIS Fraud Detection dataset was used, containing over 590,000 real-world transactions with severe class imbalance. To address this imbalance, SMOTE was applied during training.

Feature selection was performed using SHAP values, allowing the authors to reduce dimensionality while maintaining performance. The stacking ensemble significantly outperformed individual models, achieving approximately 99% accuracy and a high AUC score. To enhance transparency, the study employed SHAP, LIME, Partial Dependence Plots, and Permutation Feature Importance to explain predictions. While the framework demonstrated excellent performance, it required high computational resources and did not fully address real-time scalability. Nevertheless, this paper serves as a strong foundation for explainable and high-performance fraud detection systems.

C. Sai et al. (2023) – XAI-Driven Financial Transaction Fraud Detection

Sai et al. (2023) explored the integration of Explainable AI techniques with both machine learning and deep neural networks for financial transaction fraud detection. The study emphasized that high-performing models are often deployed without sufficient interpretability, which can lead to trust and compliance issues. The authors experimented with multiple classification models and evaluated their performance using accuracy, precision, recall, and F1-score.

Explainability techniques such as SHAP and LIME were used to provide insights into model behavior. The study showed that explainable predictions improved analysts' confidence and helped reduce false positives. However, the work lacked an in-depth comparison between ensemble learning and deep learning approaches. Additionally, the experiments were conducted on limited datasets, which raises concerns regarding generalizability. The study highlights the importance of combining performance with interpretability but leaves room for further exploration using large-scale real-world datasets.

D. Suriya and Sireesha (2025) – Credit Card Fraud Detection using XAI

Suriya and Sireesha (2025) proposed an explainable machine learning framework for detecting credit card fraud, focusing on scalability and efficient feature selection. The authors implemented multiple ML models and applied explainability techniques to interpret predictions. Their work emphasized that fraud analysts require clear explanations to justify automated decisions, especially in high-risk financial environments.

The dataset used in the study exhibited class imbalance, which was addressed using resampling techniques. The results showed improved detection accuracy and better interpretability compared to traditional models. However, the framework faced challenges in processing large transaction volumes in real time. Additionally, explanations generated by some XAI techniques were found to be complex for non-technical users. This study reinforces the need for user-friendly explanations and scalable solutions.

E. Ojo and Tomy (2025) – XAI for Credit Card Fraud Detection

Ojo and Tomy (2025) focused on improving transparency and trust in AI-based fraud detection systems. Their work emphasized regulatory compliance and ethical considerations in financial decision-making. The authors applied explainable machine learning models and evaluated their effectiveness on multiple fraud datasets.

The study demonstrated that explainable predictions improved analyst trust and reduced unnecessary transaction blocks. However, the evaluation was limited to a small number of datasets, which restricted the diversity of fraud patterns analyzed. The paper strongly advocates for explainable systems but highlights the lack of standardized evaluation metrics for explainability.

F. Yeo et al. (2025) – A Comprehensive Review on Financial Explainable AI

Yeo et al. (2025) presented a comprehensive and structured review of Explainable Artificial Intelligence techniques used in financial applications, with a strong emphasis on fraud detection, credit scoring, and risk assessment. The authors analyzed a wide range of XAI methods, including model-agnostic techniques such as LIME and SHAP, as well as model-specific explanation approaches designed for tree-based and neural network models. The study highlighted that while financial institutions increasingly rely on AI-driven systems, explainability is no longer optional due to regulatory requirements and ethical concerns.

The paper categorized XAI techniques based on global and local explanations and discussed their suitability for different financial tasks. It also examined the role of explainability in improving trust among analysts, auditors, and customers. One important contribution of this study was its identification of key challenges, such as the lack of standardized metrics for evaluating explainability and the difficulty of balancing accuracy with interpretability. The authors emphasized that many existing studies focus heavily on accuracy while treating explainability as a secondary feature.

Although the paper provided an extensive overview of existing techniques, it remained largely theoretical and did not include experimental validation on real-world datasets. Nevertheless, this review is highly valuable as it clearly identifies research gaps and establishes the need for practical, scalable, and standardized XAI frameworks in financial fraud detection systems.

G. Černevičienė and Kabašinskas (2024) – Explainable AI in Finance: A Systematic Review

Černevičienė and Kabašinskas (2024) conducted a systematic literature review focused on the application of Explainable AI within financial systems, including fraud detection, credit evaluation, and financial forecasting. The authors followed a rigorous review methodology, selecting and analyzing a large number of peer-reviewed studies to understand how explainability is currently integrated into financial AI models. Their work highlighted that many financial institutions prioritize predictive accuracy, often at the cost of transparency and interpretability.

The study discussed various explainability approaches and emphasized the importance of domain-specific explanations in finance. According to the authors, generic explanation techniques may not always meet the needs of financial analysts or regulators. The review also examined the trade-off between model complexity and interpretability, noting that highly complex models often produce explanations that are difficult for non-technical users to understand.

A key contribution of this paper was its focus on regulatory and ethical considerations. The authors emphasized that explainable systems are essential for compliance with financial regulations and for maintaining public trust. However, the review did not explore real-time deployment challenges or computational constraints. Overall, this work provides a strong theoretical foundation and supports the growing demand for explainable and responsible AI systems in financial fraud detection.

H. Prabhudesai et al. (2025) – Explainable and Responsible AI in Credit Card Fraud Detection

Prabhudesai et al. (2025) examined the role of explainable and responsible AI in credit card fraud detection systems. The study focused on ensuring fairness, transparency, and accountability in automated fraud detection models. The authors argued that financial AI systems must not only detect fraud accurately but also justify their decisions to regulators and customers. The study analyzed several machine learning models and incorporated XAI techniques to interpret predictions.

The authors highlighted that explainability improves analyst confidence and reduces disputes arising from incorrect fraud flags. The study also discussed ethical concerns such as bias in training data and the potential discrimination caused by automated systems. One of the strengths of this work was its discussion on responsible AI principles and how they can be integrated into fraud detection workflows. However, the study noted that incorporating explainability often increases system complexity and computational overhead. Additionally, the explanations generated by some techniques were found to be too technical for business users. Despite these limitations, this research provides valuable insights into how explainability and ethics can be jointly addressed in fraud detection systems, making it highly relevant for real-world financial applications.

I. Aljunaid et al. (2025) – XAI-Driven Federated Learning for Financial Fraud Detection

Aljunaid et al. (2025) proposed an innovative approach that combines federated learning with Explainable AI for financial fraud detection. The main objective of the study was to preserve data privacy while maintaining transparency and interpretability. Federated learning allows multiple financial institutions to collaboratively train models without sharing raw data, which is crucial for sensitive financial information.

The authors integrated XAI techniques to explain model predictions and ensure that decisions made by federated models remain understandable. The study demonstrated that combining federated learning with explainability can improve trust while complying with data protection regulations. Experimental results showed competitive accuracy compared to centralized models.

However, the paper also identified challenges related to communication overhead and system complexity. Federated learning requires frequent model updates across institutions, which can increase latency. Additionally, generating explanations in a distributed environment adds computational cost. Despite these challenges, the study is significant as it addresses both privacy and explainability—two critical requirements in modern financial fraud detection systems.

J. Faruk et al. (2025) – Trust and Transparency in XAI-Based Fraud Detection Systems

Faruk et al. (2025) focused on trust and transparency as core requirements for AI-based fraud detection systems. The authors argued that explainability must be designed for human understanding rather than technical completeness. The study examined several XAI techniques and evaluated how different stakeholders, including fraud analysts and managers, perceive explanations.

The findings revealed that overly complex explanations reduce usability and decision-making efficiency. The authors emphasized the need for simple, concise, and visually intuitive explanations. The study also highlighted that trust in AI systems increases significantly when users can understand the reasoning behind predictions.

While the paper provided valuable insights into human–AI interaction, it did not propose a specific fraud detection model. Instead, it focused on explanation quality and usability. This work is important as it shifts the focus from purely technical performance to user-centered explainability, which is essential for real-world adoption.

K. Chen et al. (2025) – Deep Learning in Financial Fraud Detection

Chen et al. (2025) reviewed deep learning techniques applied to financial fraud detection, including neural networks, LSTM models, and hybrid architectures. The authors highlighted that deep learning models outperform traditional machine learning approaches in detecting complex and sequential fraud patterns. However, they emphasized that these models often function as black boxes.

The study discussed the limitations of deep learning models in terms of interpretability and regulatory compliance. The authors strongly advocated for the integration of XAI techniques to make deep learning models more transparent. Although the paper provided a detailed overview of deep learning advancements, it lacked experimental validation with explainability methods.

L. Gaav et al. (2025) – Recent Advances in Credit Card Fraud Detection

Gaav et al. (2025) presented an analytical review of recent advancements in credit card fraud detection. The study examined traditional, machine learning, and ensemble-based approaches. The authors identified ensemble learning as the most effective strategy for handling class imbalance and improving detection accuracy.

However, the review revealed that explainability is often overlooked. The authors concluded that future research must integrate XAI techniques to improve transparency. This study clearly supports the need for explainable ensemble-based fraud detection systems.

M. Bhattacharyya et al. (2011) – Data Mining for Credit Card Fraud Detection

Bhattacharyya et al. (2011) conducted one of the earliest comparative studies on credit card fraud detection using data mining techniques. The study demonstrated that ensemble models outperform single classifiers. This work laid the foundation for modern ensemble-based fraud detection research.

N. Dal Pozzolo et al. (2014) – Lessons from Real-World Fraud Detection

Dal Pozzolo et al. (2014) focused on practical challenges in real-world fraud detection, such as concept drift and class imbalance. The authors emphasized adaptive learning systems. Their work remains highly relevant for modern fraud detection systems.

O. Lundberg and Lee (2017) – A Unified Approach to Interpreting Model Predictions

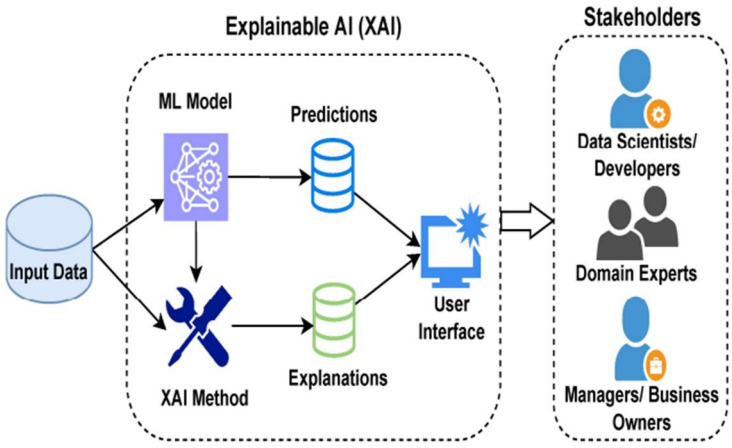
Lundberg and Lee (2017) introduced SHAP, a unified explainability framework based on game theory. SHAP became a cornerstone of Explainable AI and is widely used in financial fraud detection due to its consistency and interpretability.

III. METHODOLOGY

The methodology followed in this study focuses on building an accurate, reliable, and explainable financial fraud detection framework by integrating stacking ensemble learning with Explainable Artificial Intelligence (XAI) techniques. The overall approach is designed to address key challenges in fraud detection, such as class imbalance, evolving fraud patterns, lack of transparency, and regulatory requirements.

The process begins with data collection from real-world financial transaction datasets. These datasets typically contain a large number of transactions with diverse features related to transaction amount, time, card information, device details, and user behavior. Since fraudulent transactions represent only a small fraction of total transactions, the data is highly imbalanced. Therefore, data preprocessing plays a crucial role in improving model performance. Preprocessing steps include handling missing values, encoding categorical variables, removing irrelevant identifiers, and normalizing numerical features to ensure consistency.

To address class imbalance, resampling techniques such as the Synthetic Minority Over-sampling Technique (SMOTE) are applied to the training data. This helps the model learn meaningful fraud patterns without being biased toward legitimate transactions. After preprocessing, feature selection is performed using explainability-based methods, particularly SHAP values. SHAP helps identify the most influential features contributing to fraud predictions, reducing dimensionality while preserving important information.



The core of the methodology involves developing a stacking ensemble model. Multiple high-performing machine learning models, such as XGBoost, LightGBM, and CatBoost, are trained as base learners. These models are selected due to their strong ability to capture complex and non-linear relationships in transaction data. The predictions generated by the base models are then passed to a meta-learner, which learns how to optimally combine these outputs to produce the final fraud classification. Hyperparameter optimization techniques are used to further enhance model performance and stability.

To ensure reliability, the model is evaluated using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. Cross-validation is applied to verify generalization and prevent overfitting. Finally, Explainable AI techniques such as SHAP, LIME, permutation feature importance, and partial dependence plots are used to interpret model decisions. These explanations provide both global insights into overall model behaviour and local explanations for individual transactions. By combining high detection accuracy with transparent explanations, the proposed methodology supports trustworthy, ethical, and regulation-compliant financial fraud detection systems.

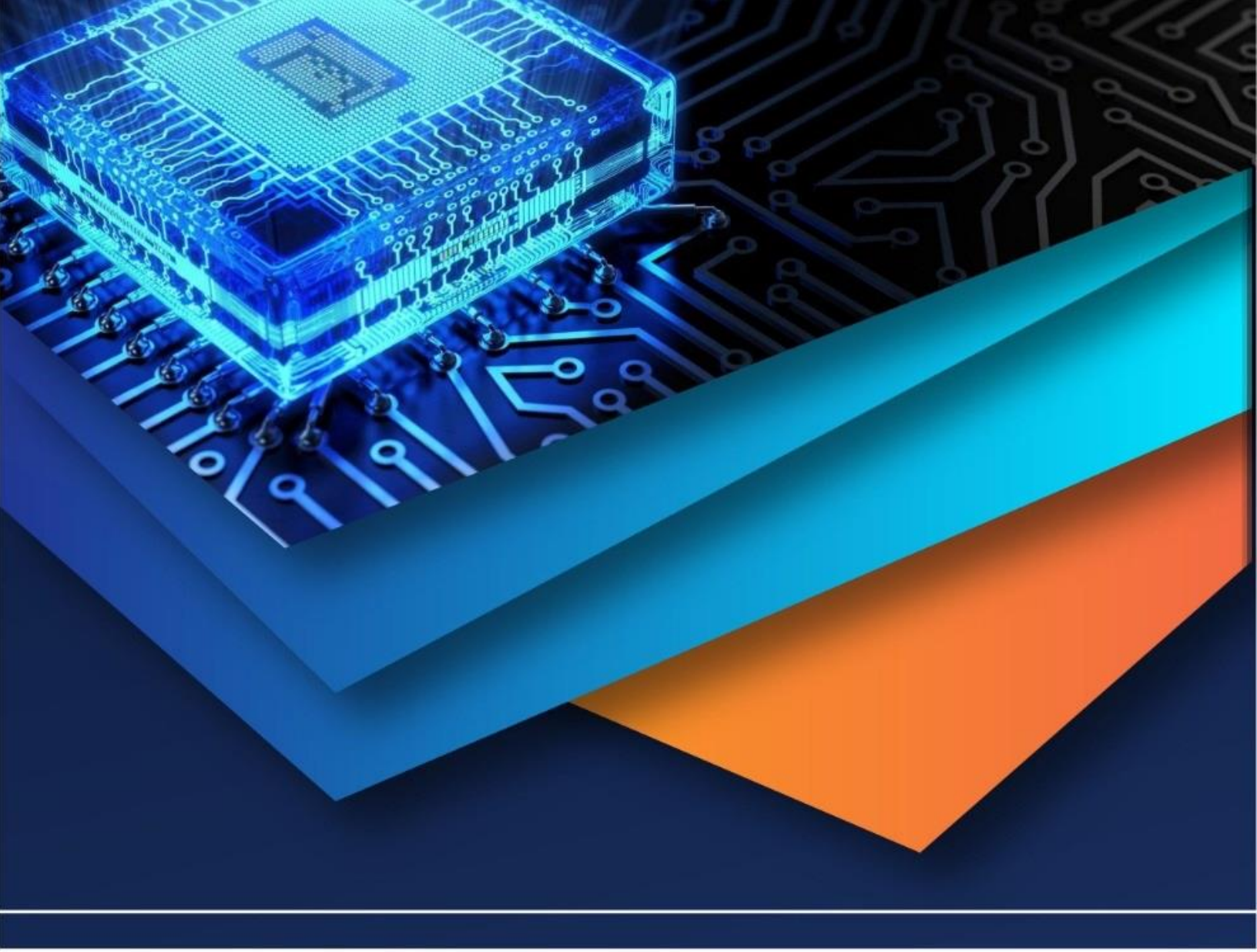
IV. CONCLUSION

Financial fraud continues to pose a significant challenge in the rapidly expanding digital financial ecosystem. With the increasing volume and complexity of online transactions, traditional fraud detection methods are no longer sufficient to effectively identify evolving fraud patterns. This review highlights the growing adoption of machine learning and ensemble-based approaches for financial fraud detection, which have demonstrated strong performance in identifying fraudulent activities. However, the lack of transparency and interpretability in many advanced models remains a critical concern, particularly in regulated financial environments. The integration of Explainable Artificial Intelligence (XAI) with stacking ensemble learning offers a promising solution to this challenge. By combining multiple high-performing models, stacking ensembles improve detection accuracy and robustness, while XAI techniques such as SHAP and LIME provide meaningful insights into model decisions. These explanations help analysts understand why transactions are classified as fraudulent or legitimate, improving trust, accountability, and regulatory compliance. This study concludes that explainability is not merely an additional feature but a fundamental requirement for modern fraud detection systems. Future research should focus on improving real-time scalability, developing user-friendly explanation methods, and ensuring ethical and responsible AI deployment. Overall, the combination of accuracy and interpretability is essential for building reliable, transparent, and trustworthy financial fraud detection systems.

REFERENCES

[1] Fahad Almalki and Mehedi Masud, “Financial Fraud Detection Using Explainable AI and Stacking Ensemble Methods,” arXiv preprint arXiv:2505.10050, May 2025.
[2] Hasan, M., Rahman, S., and Hossain, M., “Explainable Artificial Intelligence in Credit Card Fraud Detection,” Journal of Computer Science and Technology Studies, vol. 6, no. 2, pp. 45–58, 2024.
[3] Sai, C. V., Das, D., Elmitwally, N., Elezaj, O., and Islam, M. B., “Explainable AI-Driven Financial Transaction Fraud Detection Using Machine Learning and Deep Neural Networks,” SSRN Preprint, 2023. DOI: 10.2139/ssrn.4439980
[4] Suriya, R., and Sireesha, M., “Credit Card Fraud Detection Using Explainable Artificial Intelligence,” Journal of Information Systems Engineering and Management, vol. 10, no. 1, 2025.
[5] Ojo, A., and Tomy, K., “Explainable Artificial Intelligence for Credit Card Fraud Detection,” World Journal of Advanced Research and Reviews, vol. 15, no. 2, pp. 112–121, 2025.
[6] Yeo, K., Lim, S., and Tan, W., “A Comprehensive Review on Financial Explainable Artificial Intelligence,” Artificial Intelligence Review, vol. 58, no. 4, pp. 1–29, 2025.
[7] Černevičienė, J., and Kabašinskas, A., “Explainable AI in Finance: A Systematic Review,” Artificial Intelligence Review, vol. 57, no. 3, pp. 345–372, 2024.
[8] Prabhudesai, S., Kulkarni, R., and Patil, A., “Explainable and Responsible AI in Credit Card Fraud Detection,” University of Mumbai (SAKEC), Technical Report, 2025.
[9] Aljunaid, S., Alshamrani, A., and Khan, M., “Explainable AI-Driven Federated Learning for Financial Fraud Detection,” Journal of Risk and Financial Management, vol. 18, no. 1, 2025. DOI: 10.3390/jrfm18010045
[10] Faruk, M., Rahman, T., and Hossain, A., “Explainable AI for Fraud Detection: Trust and Transparency,” Financial Security Systems Research Paper, 2025.
[11] Chen, Y., Li, Z., and Wang, X., “Deep Learning in Financial Fraud Detection: Innovations and Applications,” Data Science and Management, Elsevier, vol. 7, pp. 88–102, 2025.
[12] Gaav, A., Mehta, P., and Shah, N., “Recent Advances in Credit Card Fraud Detection: An Analytical Review,” Journal of Future AI and Technologies, vol. 4, no. 1, pp. 25–39, 2025.
[13] Bhattacharyya, S., Jha, S., Tharakunnel, K., and Westland, J. C., “Data Mining for Credit Card Fraud: A Comparative Study,” Decision Support Systems, vol. 50, no. 3, pp. 602–613, DOI: 10.1016/j.dss.2010.08.008
[14] Dal Pozzolo, A., Caelen, O., Le Borgne, Y. A., Waterschoot, S., and Bontempi, G., “Learned Lessons in Credit Card Fraud Detection from a Practitioner Perspective,” Expert Systems with Applications, vol. 41, no. 10, pp. 4915–4928, 2014. DOI: 10.1016/j.eswa.2014.02.026
[15] Lundberg, S. M., and Lee, S. I., “A Unified Approach to Interpreting Model Predictions,” Advances in Neural Information Processing Systems (NeurIPS), pp. 4765–4774, 2017.

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)