INTERNATIONAL JOURNAL OF COMPUTER SCIENCE
Scholorly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
Since 2012
www.ijcsjournal.com Volume 14, Issue 1, No 26 2026 ISSN: 2348-6600
REFERENCE ID: IJCS-699 PAGE NO: 001-005

AN EXPLAINABLE AI FRAMEWORK FOR INTERPRETABLE AND RELIABLE FRAUD DETECTION IN FINANCIAL SYSTEMS

R. Sangavi
Scholar,
Department of Computer Application,
A.V.V.M. Sri Pushpam College (Autonomous),
Tiruvallur, Tamil Nadu, India.

Dr. D. Ragupathi
Head of the Department,
Department of Computer Applications,
A.V.V.M. Sri Pushpam College (Autonomous),
Tiruvallur, Tamil Nadu, India.

Abstract

Fraud detection has become a critical challenge in modern financial systems due to the rapid growth of digital transactions. Although machine learning models have demonstrated high accuracy in detecting fraudulent activities, their black-box nature limits transparency and hinders trust in real-world applications. This paper proposes an Explainable Artificial Intelligence (XAI)-based framework to enhance the interpretability of fraud detection models without compromising predictive performance. The proposed approach integrates advanced machine learning classifiers with post-hoc explanation techniques, including Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive explanations (SHAP), to generate both local and global insights into model decisions. Experimental evaluation on real-world datasets shows that the framework achieves high detection accuracy while significantly improving transparency, trustworthiness, and regulatory compliance. The results demonstrate that XAI plays a vital role in bridging the gap between model performance and human interpretability in fraud detection systems.

Keywords: Explainable Artificial Intelligence (XAI), Fraud Detection, Machine Learning, Model Interpretability, SHAP, LIME, Financial Security.

Literature Review:

Recent advancements in fraud detection have increasingly focused on integrating Explainable Artificial Intelligence (XAI) techniques to address the limitations of traditional black-box machine learning models. Several studies have contributed significantly to this domain.

Adadi and Berrada [1] presented a comprehensive survey of Explainable Artificial Intelligence, categorizing various XAI methods into intrinsic and post-hoc approaches. Their work emphasized the growing need for interpretability in high-stakes domains such as finance, healthcare, and law enforcement, where transparency and accountability are critical.

INTERNATIONAL JOURNAL OF COMPUTER SCIENCE
Scholorly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
Since 2012
www.ijcsjournal.com Volume 14, Issue 1, No 25 2026 ISSN: 2348-6600
REFERENCE ID: IJCS-695 PAGE NO: 094-098

Ribeiro et al. [2] introduced Local Interpretable Model-Agnostic Explanations (LIME), a technique designed to explain individual predictions of any machine learning model. LIME approximates complex models locally using simpler, interpretable models, thereby enabling users to understand the reasoning behind specific predictions. This approach is particularly useful in fraud detection, where understanding individual transaction decisions is essential.

Lundberg and Lee [3] proposed SHapley Additive exPlanations (SHAP), a unified framework based on cooperative game theory. SHAP assigns importance values to each feature by fairly distributing their contribution to the prediction. Unlike other methods, SHAP provides both local and global interpretability, making it highly effective for analyzing fraud detection models and identifying key risk factors.

Molnar [4] provided a practical perspective on interpretable machine learning, discussing various techniques such as feature importance, partial dependence plots, and surrogate models. The work highlights how interpretability methods can be applied in real-world scenarios, including fraud detection, to improve model transparency and reduce false positives.

Doshi-Velez and Kim [5] emphasized the need for a rigorous framework to evaluate interpretability in machine learning systems. They proposed different evaluation strategies, including application-grounded and human-centered approaches, ensuring that explanations are meaningful and aligned with user requirements.

Overall, these studies highlight the importance of balancing predictive accuracy with interpretability in fraud detection systems. The integration of XAI techniques enables better decision-making, enhances trust, and ensures compliance with regulatory standards.

Problem Statement

Traditional fraud detection systems predominantly rely on advanced machine learning and deep learning models to identify fraudulent activities. While these models achieve high predictive accuracy, they often operate as black-box systems, making their decision-making processes difficult to interpret.

The lack of transparency in these systems creates significant challenges in understanding how specific predictions are made. This limitation hinders the ability of analysts and stakeholders to validate model decisions and investigate flagged transactions effectively. Furthermore, the presence of high false positive rates leads to inconvenience for legitimate users and increases the operational burden on financial institutions.

In addition, regulatory frameworks in financial systems require clear justification for automated decisions. The inability of traditional models to provide interpretable explanations makes compliance with such regulations difficult. As a result, these limitations reduce trust and confidence among

INTERNATIONAL JOURNAL OF COMPUTER SCIENCE
Scholorly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
Since 2012
www.ijcsjournal.com Volume 14, Issue 1, No 25 2026 ISSN: 2348-6600
REFERENCE ID: IJCS-695 PAGE NO: 094-098

stakeholders, including customers, auditors, and financial analysts.

Therefore, there is a critical need for a fraud detection framework that not only delivers high accuracy but also ensures transparency, interpretability, and trustworthiness in decision-making processes.

Proposed Methodology

System Overview

The proposed system aims to enhance fraud detection by integrating machine learning models with Explainable Artificial Intelligence (XAI) techniques. While conventional models provide high predictive accuracy, they often lack interpretability. To address this limitation, the proposed framework incorporates post-hoc explanation methods that enable both local and global interpretability of model predictions. This integration allows stakeholders to understand the reasoning behind each classification, thereby improving transparency, trust, and decision-making in fraud detection systems.

Methodology Workflow

The overall workflow of the proposed system consists of several key stages, including data processing, model training, prediction, and explanation generation. Initially, transaction data is collected from relevant financial datasets, which may include attributes such as transaction amount, time, location, and user behavior patterns. The collected data is then preprocessed through cleaning, normalization, and handling of missing values to ensure data quality and consistency.

Following preprocessing, machine learning models such as Random Forest, Extreme Gradient Boosting (XGBoost), and Neural Networks are trained to classify transactions as fraudulent or non-fraudulent. These models are selected due to their strong performance in handling complex and high-dimensional data.

Once the model is trained, it is used to predict whether a given transaction is fraudulent or legitimate. To improve interpretability, XAI techniques are applied to generate explanations for each prediction. Local Interpretable Model-Agnostic Explanations (LIME) are used to provide instance-level explanations, highlighting the contribution of features for individual predictions. Additionally, SHapley Additive exPlanations (SHAP) are employed to provide both local and global insights, identifying overall feature importance and model behavior.

System Architecture

The proposed system integrates machine learning models with Explainable Artificial Intelligence (XAI) techniques to improve the interpretability of fraud detection. It provides accurate predictions along with explanations, enhancing transparency and trust in decision-making.

- **Customer:** Initiates financial transactions that are monitored by the system.

INTERNATIONAL JOURNAL OF COMPUTER SCIENCE
Scholorly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
Since 2012
www.ijcsjournal.com Volume 14, Issue 1, No 25 2026 ISSN: 2348-6600
REFERENCE ID: IJCS-695 PAGE NO: 094-098

- **Financial Institution:** Collects and processes transaction data for fraud analysis.
- **XAI-based Fraud Detection Engine:** Core component that performs classification using machine learning models and generates interpretable explanations using XAI techniques.
- **Fraud Analyst:** Reviews flagged transactions and utilizes explanations to validate decisions and investigate potential fraud cases.
- **Regulatory Authority:** Ensures that the system complies with financial regulations by requiring transparent and explainable decision-making processes.

This architecture ensures a seamless flow of data and information between stakeholders while maintaining transparency and accountability in fraud detection.

System Analysis

Existing System

Traditional fraud detection systems rely on black-box machine learning models that achieve high accuracy but lack interpretability. This limits transparency, makes decision auditing difficult, and reduces trust among stakeholders. Additionally, these systems struggle to meet regulatory requirements due to the absence of clear explanations.

Proposed System

The proposed system integrates Explainable Artificial Intelligence (XAI) techniques with machine learning models to provide interpretable predictions. It enhances transparency, improves trust, supports regulatory compliance, and reduces false positives by offering clear explanations for model decisions.

Results and Discussion

The experimental results demonstrate that the proposed system achieves high fraud detection accuracy while effectively reducing false positives. Compared to traditional black-box models, the integration of Explainable Artificial Intelligence (XAI) techniques significantly improves interpretability and transparency in decision-making.

Performance evaluation metrics such as accuracy, precision, recall, and F1-score indicate that the model maintains strong predictive capability. In addition, XAI methods such as SHAP and LIME provide meaningful insights into feature contributions, enabling both local and global understanding of model behavior.

The results confirm that the system not only delivers reliable predictions but also offers clear explanations for each decision. This improves trust among stakeholders and supports better validation of fraud detection outcomes, making the proposed approach more practical for real-world financial applications.

INTERNATIONAL JOURNAL OF COMPUTER SCIENCE
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
Since 2012
www.ijcsjournal.com Volume 14, Issue 1, No 25 2026 ISSN: 2348-6600
REFERENCE ID: IJCS-695 PAGE NO: 094-098

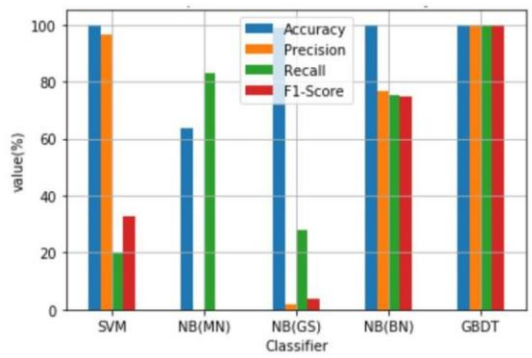

Fig 1: Performance Evaluation of Models

Conclusion

This study demonstrates that integrating Explainable Artificial Intelligence (XAI) techniques into fraud detection systems significantly improves transparency while maintaining high predictive accuracy. The proposed approach enhances interpretability, builds trust among stakeholders, and supports regulatory compliance. By providing clear explanations for model decisions, the system enables more effective and reliable fraud detection in financial applications.

Future Work

Future work will focus on extending the proposed system to real-time fraud detection environments. Additionally, the integration of deep learning models with advanced explainability techniques can be explored. Developing interactive visualization dashboards and investigating counterfactual explanations are also potential directions to further enhance interpretability and user understanding.

References

1. A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, 2018.
2. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), San Francisco, CA, USA, 2016, pp. 1135–1144.
3. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. 31st Int. Conf. Neural Information Processing Systems (NeurIPS), 2017, pp. 4765–4774.
4. C. Molnar, *Interpretable Machine Learning*, 2nd ed. Leanpub, 2022.
5. F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.