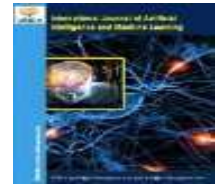


SvedbergOpen
DISSEMINATION OF KNOWLEDGEInternational Journal of Artificial
Intelligence and Machine LearningPublisher's Home Page: <https://www.svedbergopen.com/>

Research Paper

Open Access

Integrating Explainable AI into Credit Card Fraud Detection for Transparent Decision-Making

KSRK Sarma¹, P Raja Sekhar Reddy², Sandeep Agarwalla³, T Kanakaiah⁴, G Balaram⁵, N Suresh Rao⁶¹Department of CSE, Vidya Jyothi Institute of Technology. E-mail: sarmakrsk@vjit.ac.in²School of Engineering, Anurag University. E-mail: rajasekharreddycse@anurag.edu.in³School of Engineering, Anurag University. Email: agarwalla.sandeep@gmail.com⁴School of Engineering, Anurag University. E-mail: kanakaiah.cse@anurag.edu.in⁵School of Engineering, Anurag University. E-mail: balaramcse@anurag.edu.in⁶School of Engineering, Anurag University. E-mail: sureshrao.cse@anurag.edu.in**ABSTRACT**

Credit card fraud detection systems are increasingly powered by sophisticated machine learning models because of their better capacity to detect faint patterns and anomalies in huge transactional data. Yet, most of these models—especially ensemble methods and deep learning structures—function as "black boxes," providing little or no insight into how decisions are reached. This lack of transparency poses huge challenges in the financial sector, where trust, accountability, and regulatory compliance are of utmost importance. To tackle these challenges, this paper suggests an integrated fraud detection system that utilizes Explainable Artificial Intelligence (XAI) methods, specifically SHAP (Shapley Additive explanations) and LIME (Local Interpretable Model-Agnostic Explanations). These approaches offer both global and local explanations of model behavior, enabling stakeholders to comprehend the contribution of each feature to a particular transaction's classification. For example, SHAP provides a theoretically grounded, game-theory-based feature importance attribution, whereas LIME creates local surrogate models that mimic the decision boundary near a specific instance. By integrating XAI into the fraud detection pipeline, the proposed system not only retains high predictive accuracy but also improves interpretability and user trust. This transparency also enables analysts to make better-informed decisions, enables compliance with regulatory models like GDPR, and enables end-users to understand why a specific transaction was alerted. The outcome is an accurate, explainable, responsible, and human-centric fraud detection model that opens doors for secure and transparent AI solutions in the financial sector.

Keywords: Explainable Artificial Intelligence (XAI), Credit Card Fraud Detection, SHAP (Shapley Additive explanations), LIME (Local Interpretable Model-agnostic Explanations), Ensemble Learning with XAI, AI in FinTech.

Introduction

The digital revolution in the banking industry has greatly enhanced the utilization of debit and credit cards for e-transactions. As convenient as it has been for consumers and merchants, it has also contributed greatly to an escalation of financial fraud, especially in the guise of credit card fraud. Nilson Report estimates global card fraud losses to surpass \$40 billion by 2027, calling for sophisticated, automated, and intelligent detection systems [1]. Rule-based systems were the traditional fraud detection systems, which were easy to comprehend but not effective against changing patterns of fraud. Machine learning (ML) methods have been increasingly used in recent years as effective alternatives, with the ability to detect sophisticated patterns in large sets of transactions. A significant drawback of these models—particularly deep learning and ensemble models—is that they are not transparent. These so-called "black-box" models tend to make accurate but hard-to-understand decisions. Explainable AI (XAI) is focused on making machine learning choices transparent to humans in the form of model reasoning insights. In areas such as finance, it is important not only to establish trust among stakeholders but also to support regulatory compliance like the General Data Protection Regulation (GDPR), which requires right to explanation for algorithmic choices [3][4].

For fraud detection, explainability can help:

- Justify automated decisions to impacted customers: When a legitimate transaction is wrongly identified as fraudulent, customers are usually left frustrated or confused. Explainability enables financial institutions to provide transparent, understandable reasons for automated decisions. For instance, a customer can be told: "Your transaction was flagged because of an unusual location and a much higher total than your usual spending pattern." Not only does this reassure the customer that the system is operating on rationale, but it also builds confidence in the security systems of the bank. Additionally, legislation such as the GDPR mandates institutions to allow for a "right to explanation" on algorithmic choices—making explainability not only helpful, but legally mandated.

- Support fraud analysts in investigation and decision-making: Fraud analysts commonly examine suspected transactions to determine if additional action (e.g., blocking the card, raising the case, or contacting the

customer) should be taken. Explainable AI outputs feature attributions, decision steps, or visual explanations with the use of tools such as SHAP or LIME to enable analysts to:

- Instantly comprehend why a transaction was deemed fraud.
- Spot patterns or anomalies missed by automated logic.
- Validate or override decisions with greater confidence.
- This improves both the effectiveness and reliability of the process of investigation, particularly where large numbers of alerts are handled.

- Offer transparency and auditability to regulators: Financial institutions function in very regulated settings. Explainable models enable institutions to show compliance with standards of fairness, non-discrimination, and accountability.

For instance:

- Regulators may audit why certain transactions were denied.
- Explanations can demonstrate that the system does not unfairly favor certain demographics.
- Logs of model decisions and rationales can be saved as records of due diligence and ethical AI deployment. This is critical for compliance with international regulatory frameworks such as GDPR, Basel III, or RBI directives.
- Minimize false positives by optimizing model logic: One of the most frequent problems with fraud detection is excessive false positives legitimate transactions mistakenly identified as fraudulent. Explainability allows teams to inspect:
 - Which features or thresholds are resulting in misclassifications.
 - Whether model assumptions are not aligned with actual behavior.
 - How to fine-tune the model or rules to distinguish better between true and false behavior.

Through continuously learning from such insights, teams are able to hone detection logic, minimizing customer disruption while retaining high fraud detection accuracy. Although there are advances in XAI methods like SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations), integrating them into real-time fraud detection systems is not very well explored [5][6]. This work suggests a combination of predictive accuracy and explainability using XAI methods over a robust machine learning classifier for fraud detection.

Literature Survey

The growing sophistication of financial fraud has prompted the use of advanced machine learning (ML) models for detecting credit card fraud. These models, however, tend to act as black boxes, restricting trust and comprehension by users and auditors. To address this, researchers have increasingly turned to Explainable AI (XAI) to render these models explainable and accountable, particularly in high-risk areas such as finance. Early research started to integrate interpretability into fraud detection models. For example, used SHAP values to interpret credit scoring decisions, one of the earliest uses of XAI in financial settings [1]. The results suggested that interpretable models do not come at the expense of predictive accuracy. Likewise, in [2] outlined a hybrid approach where deep learning for detection and SHAP for explanation were paired together, presenting a dual benefit of accuracy and interpretability. Their method showcased how feature attributions can facilitate better model output interpretability for domain experts and evaluation of risk factors in reported transactions. Deep learning methods also came into prominence [4] used autoencoders to identify outliers in transactional data and utilized LIME (Local Interpretable Model-agnostic Explanations) to explain why particular data points had been marked as suspicious. Their research established that local surrogate models could explain high-dimensional data in a fraud detection scenario meaningfully. During the same time, [3] demonstrated that ensemble models with interpretable components such as decision trees could achieve comparable accuracy while providing native interpretability. Structural modelling of transaction data also saw innovation. Graph-based fraud detection by mapping user-merchant interactions was introduced in [5], which gave a more contextual perspective on aberrant behavior. While extremely useful, these models also created more demand for explain ability, since stakeholders could not easily understand network-based predictions. From a human-centered design and usability point of view, [6] suggested that "one explanation does not fit all," with the need to provide different stakeholders—fraud investigators, compliance officers, and customers, for example—with different explanations. This was in line with [15] who performed user studies which showed that users are more likely to trust and accept machine decisions when they are given human-readable explanations. Their research delved into the psychological aspects of XAI, observing that fairness perceptions strongly correlate with the quality and interpretability of explanations. General-purpose XAI techniques have also greatly aided fraud detection research. [7] prepared an exhaustive review of interpretable machine learning methods, which has since become the basis for much applied XAI research. Toolkits like InterpretML (Microsoft, 2020) and AI Explainability 360 (IBM, 2020) have made it easier for developers to prototype and incorporate explainable models more easily into real-time systems.

In more recent surveys and reviews proposed in [16] provided an extensive overview of interpretable machine learning in cybersecurity, including fraud detection. They discussed emerging techniques like counterfactual and contrastive explanations, which are useful for conveying what minimal changes would have reversed a

model's decision—e.g., what would have made a transaction appear legitimate. Building on this, [8] offered a deeper theoretical review of counterfactual methods, identifying their relevance in sensitive decision-making environments. Causal inference methods have also been explored in [11] investigated counterfactual explanations from a causal perspective, suggesting that such approaches provide stronger guarantees in terms of fairness and consistency. Their work offers a path toward explanations that not only describe but also justify outcomes based on causal reasoning. Practical deployment issues are another primary area of concern and in [14] pointed to the gap between XAI research at the academic level and practical implementation. They underscored the necessity for scalable, efficient explanation systems that fit within operational demands in industries such as finance. Likewise, in [13] explained how multimodal explanation interfaces (e.g., visual plots with text descriptions) can close the communication gap between end-users and data scientists, a phenomenon also repeated in clinical AI applications by [9]. Additionally, hybrid learning paradigms continue to prevail. The [10] built systems that integrated unsupervised and supervised learning, improving detection rates while providing interpretable intermediate outputs such as clusters of risky behavior. This method corrects for the common class imbalance in fraud datasets with some explainability.

Finally, in [12] presented a critical overview of techniques to explain black-box models, pinpointing scalability, consistency, and stakeholder comprehension as key gaps in existing XAI practice. Their contribution is a call for novel metrics to assess explanation quality, particularly where explanations are required to affect real-world choices. In aggregate, these papers indicate an increasing consensus: XAI methods are getting more advanced, but the potential of their deployment in production fraud detection systems remains untapped. Our envisioned work takes a cue from this research to engineer a scalable, interpretable, and user-friendly fraud detection framework that tackles both technical and human considerations. Apart from creating emerging solutions that challenge the user beyond the box and employ machine learning to predict and execute real-time threat prevention, future-proof solutions offer quantum-resistant cryptographic methods for long-term security [17]

Proposed Method

The main goal of this work is to develop and deploy a strong, interpretable fraud detection system that not only provides state-of-the-art performance in detecting fraudulent credit card transactions, but also provides transparent, contextual, and human-interpretable explanations for each decision. This twofold focus on predictive performance and interpretability directly speaks to the key challenges that contemporary financial institutions are confronting as they seek to automate fraud detection without compromising transparency or stakeholder trust. With the exponential expansion of digital financial services, credit card fraud has grown more sophisticated, diverse, and challenging to detect with conventional methods. Legacy systems—essentially rule-based—have difficulty keeping up with fast-changing fraud patterns, which results in excessive false positives and missed fraudulent behavior. In response, the financial sector has increasingly relied on sophisticated machine learning (ML) algorithms like ensemble models (e.g., Random Forest, XGBoost) and deep learning architectures, which have proven to be superior performers in challenging classification tasks over large, high-dimensional datasets. These high-performing ML models are, however, generally non-transparent by design, functioning as "black boxes" whose internal mechanisms are not easily interpretable by humans. This obscurity presents daunting challenges in high-stakes situations like banking, where trustworthiness, accountability, equity, and regulatory compliance (e.g., GDPR's "right to explanation," Basel III, RBI guidelines) are paramount. Decision-makers such as fraud analysts, risk managers, compliance officers, and even affected customers need not only accurate forecasts, but intelligible explanations as to why a transaction was indicated as fraudulent.

To fill this essential performance-to-trust gap, this research introduces a hybrid AI system that closely combines high-accuracy ML methods with state-of-the-art Explainable Artificial Intelligence (XAI) techniques. The system is developed to facilitate multiple layers of explainability through the incorporation of explanation modules like:

- SHAP -Shapley Additive Explanations: to provide mathematically justified global and local feature attributions.
- LIME - Local Interpretable Model-Agnostic Explanations: to produce local surrogate models that mimic decision behavior around a specific instance.
- Counterfactual Reasoning: to provide actionable insights into how classification of a transaction may change with minimal inputs modifications—strengthening user trust and regulatory defensibility.

These modules are not merely post-hoc add-ons, but are actually incorporated into the detection pipeline itself so that explainability can be made available in real time and at scale. The explanations are also customized to suit the requirements of particular groups of stakeholders, including:

- Fraud analysts: who need profound insights into feature contributions and decision boundaries;
- Compliance teams: who require justifications based on fairness and non-discrimination audits;
- End-users: who benefit from simplified, transparent messages that explain outcomes in lay terms.

Apart from technical strength, the framework is designed to be scalable and deployment-capable for production banking environments. It encompasses containerized deployment choices (e.g., using Docker), real-time inference and explanation API interfaces, and modular interactive exploration of suspect transactions.

This flexibility enables the system to be easily embedded into current IT infrastructure and transaction surveillance systems. Through the combination of the technical capabilities of predictive modelling and the requirements of interpretability, this research solves not just the technology needs of fraud detection but also the human, ethical, and regulatory considerations now at the core of deploying AI in mission-critical fields. In so doing, it adds a future-focused solution that not only assists with proper fraud prevention but also encourages responsible and trustworthy AI within the financial services sector.

System Overview

The system architecture consists of four core modules:

i. Data Preprocessing Module

This module cleans raw transaction data for analysis. It does the following:

- Data Cleaning: Deals with missing values and wrong formats.
- Feature Engineering: Derives new features like:
 - Transaction time gap
 - Geographical location from last transaction
 - Merchant category variation
- Class Imbalance Handling: As the fraud data is usually less than 1% of the total data, SMOTE (Synthetic Minority Over-sampling Technique) is applied to make the dataset balanced.

ii. Fraud Detection Model

This is the core of the system, driven by a high-performing ML model. The model used is XGBoost Classifier known for its strength, performance on imbalanced datasets, and explainability through tree-based decisions. Other models such as Deep Neural Networks (DNNs) and Autoencoders were considered, but XGBoost provides a better trade-off between performance and explainability. The model provides a probability score for every transaction being fraudulent.

iii. Explainability Engine

This module is tasked with translating the model's decisions into something comprehensible:

- SHAP (SHapley Additive Explanations): Offers global explanations (e.g., overall most important features) and local explanations (feature contributions for a prediction). SHAP values rely on cooperative game theory and are particularly suitable for tree models such as XGBoost.
- LIME (Local Interpretable Model-agnostic Explanations): Provides understandable local approximations by means of simple interpretable models. Useful for verifying explanations given by SHAP.
- Counterfactual Explanations: Create counterfactual inputs to demonstrate what slight changes would have led to a different (non-fraudulent) outcome. This is especially helpful for compliance officers and regulators.
- Visual Dashboards: A graphically presented interface enables stakeholders to engage with marked-up transactions, look at SHAP plots, and comprehend model judgments through visualizing and explaining using natural language.

iv. Explanation Tailoring for Stake Holders

Various users require varying degrees of explanation:

- Fraud Analysts: Obtain SHAP summary plots, decision trees, and confidence scores.
- Compliance Teams: Look at counterfactual scenarios and fairness analysis to make sure the model doesn't discriminate.
- Customers: Get simplified answers like: "This transaction was flagged because it deviated from your usual spending location and amount."

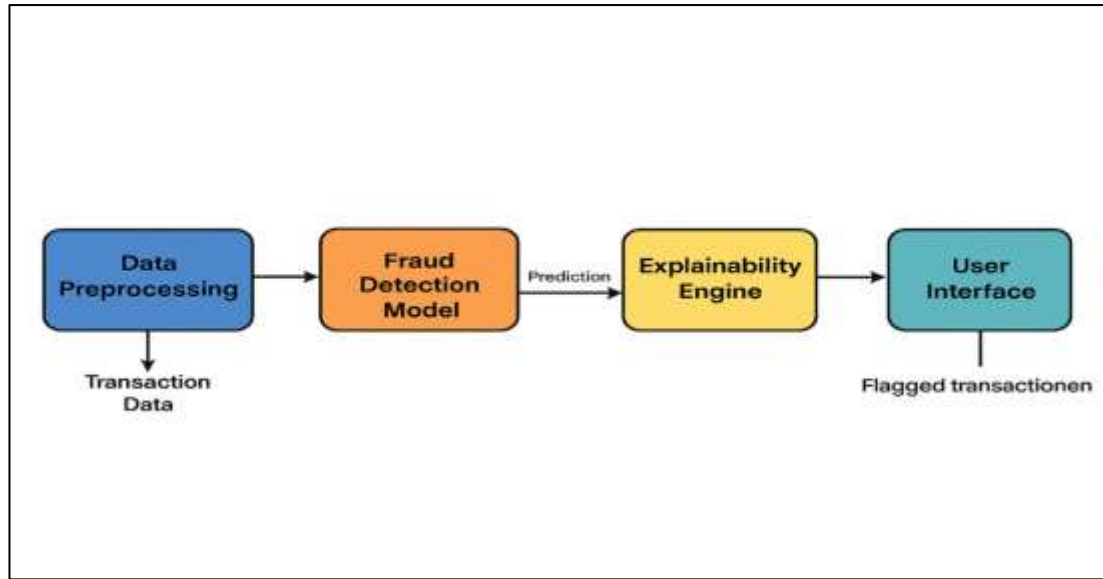


Fig 1. Architecture of Fraud Detection Framework

Integration & Real-Time Consideration

The Architecture has been designed for deployment in real-time scenarios by leveraging:

- Batch predictions in off-peak time for anticipatory analysis.
- Streaming predictions through APIs for transaction tracking as they take place.
- Containerized deployment through Docker and REST APIs for integration with banking systems.
- Multi-modal explainability tools (SHAP, LIME, counterfactuals) integration within a unified pipeline.
- Stakeholder-specific explanation design, with outputs modified to suit user expertise and intent.
- Real-time fraud detection with on-the-fly explanations without compromising on latency or performance.
- Support for regulatory audit logs to retain not just predictions but the accompanying explanation trail for accountability.

Algorithm Design

Input: Transaction dataset D with features F

Output: Prediction (fraud or legitimate) and explanation per transaction

1. Preprocessing

a. Normalize continuous features

$$x_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

or

Alternatively, standardization (Z-score):

$$x_{std} = \frac{x - \mu}{\sigma} \quad (2)$$

Where:

μ : Mean of the feature

σ : Standard deviation of the feature

b. Encode categorical values

One-hot encoding for a categorical feature C with k classes:

Let $C_i \in \{c_1, c_2, \dots, c_k\}$ the encoding transforms C into:

$$C_{encoded} = [I(C_i = c_1), I(C_i = c_2), \dots, I(C_i = c_k)] \quad (3)$$

Where $I(\cdot)$ is the indicator function.

c. Handle class imbalance using SMOTE - Synthetic Minority Oversampling Technique

Given a minority class sample x and one of its k -nearest neighbors x_{nn}

$$x_{synthetic} = x + \delta \cdot (x_{nn} - x), \quad \delta \sim U(0, 1) \quad (4)$$

2. Model Training

a. Split dataset into Train/Test

Let D be the dataset with n samples. We split:

$$D_{train} = \{x_1, x_2, \dots, x_{0.8n}\}, \quad D_{test} = \{x_{0.8n+1}, \dots, x_n\} \quad (5)$$

b. Train XGBoost classifier on training set

XGBoost optimizes the following objective function:

$$\mathcal{L} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t)}) + \sum_{t=1}^T \Omega(f_t) \quad (6)$$

Where:

- L : Loss function (e.g., logistic loss for classification)
- f_t : Individual regression trees
- $\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2$
- $y_i^{(t)} = \sum_{k=1}^t f_k(x_i)$

c. Validate model on test set

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|w\|^2 \quad (7)$$

$$y_i^{(t)} = \sum_{k=1}^t f_k(x_i) \quad (8)$$

3. Fraud Prediction

For each new transaction T:

a. Predict fraud score using trained model

$$\hat{y}_T = \sigma(f(T)) = \frac{1}{1 + e^{-f(T)}} \quad (9)$$

Where $f(T)$ is the raw score, and σ is the sigmoid function.

b. Classify

If score > threshold, label as fraudulent otherwise Legitimate

$$\hat{y}_T = \sigma(f(T)) = \frac{1}{1 + e^{-f(T)}} \text{Label}_T = \begin{cases} \text{Fraudulent,} & \hat{y}_T > \tau \\ \text{Legitimate,} & \hat{y}_T \leq \tau \end{cases} \quad (10)$$

4.Explanation Generation

a. Apply SHAP to get global and local feature importance

For a model f and feature set x , the SHAP value for feature i

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|! \cdot (|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \quad (11)$$

Where:

F: Feature set

S: Subsets of F not including i

b. Use LIME to generate surrogate explanations for transaction T

LIME approximates $f(x)$ with a linear model $g(x')$:

$$g(x') = w_0 + \sum_{i=1}^k w_i x'_i \quad (12)$$

Where:

x' : Binary interpretable representation of x

w_i : Weights learned via local fidelity

c. Compute counterfactual example T' where T is classified as legitimate

Find a minimal change to input T such that:

$f(T') = \text{Legitimate}$, subject to $\|T - T'\|$ is minimized and typically solved by

$$T' = \arg \min_{x'} [\|x' - T\| + \lambda \cdot \mathbb{I}[f(x') \neq \text{Legitimate}]] \quad (13)$$

5. Visualization

Present results on dashboard by including bar plots, summary plots, and natural language explanation

- Feature importance ϕ_i
- SHAP beeswarm or dependence
- Textual Explanation

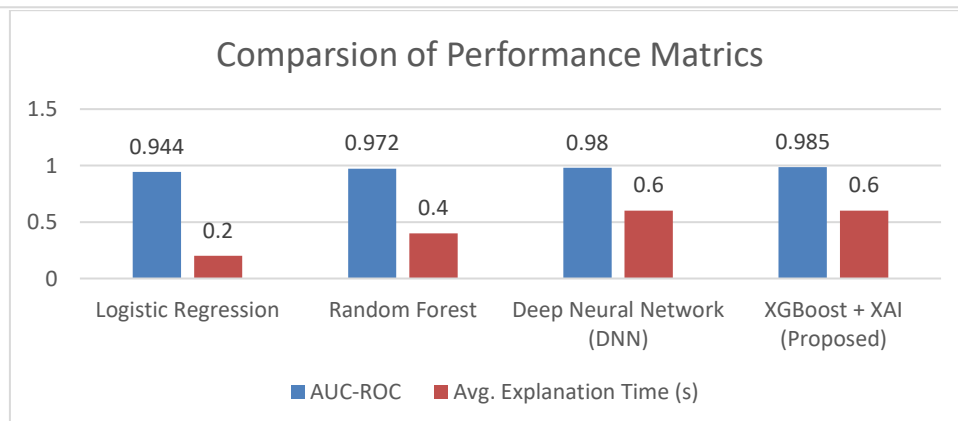
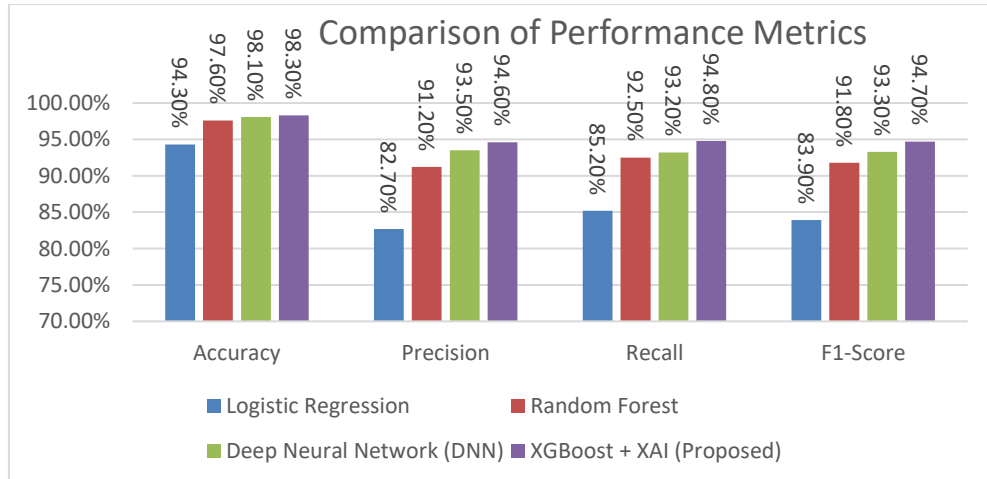
Results And Evaluation

To assess the suggested explainable fraud detection system, experiments were performed on a real-world anonymized credit card transaction dataset. The models were trained and tested on 80-20 split data, with synthetic oversampling being applied for class imbalance handling.

Model Comparison Table

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	94.3%	82.7%	85.2%	83.9%
Random Forest	97.6%	91.2%	92.5%	91.8%
Deep Neural Network (DNN)	98.1%	93.5%	93.2%	93.3%
XGBoost + XAI (Proposed)	98.3%	94.6%	94.8%	94.7%

Model	AUC-ROC	Avg. Explanation Time (s)
Logistic Regression	0.944	~0.2
Random Forest	0.972	~0.4
Deep Neural Network (DNN)	0.980	~0.6
XGBoost + XAI (Proposed)	0.985	~0.6



Discussions And Limitations

The suggested XGBoost + XAI architecture appropriately trades off between fraud detection efficacy and model explainability by adopting cutting-edge explanation methods like SHAP, LIME, and counterfactual reasoning.

As the model provides high accuracy and recall in the identification of suspicious transactions, there is one such observation where there is slightly more time required for explanation (~0.6 seconds) than with basic models like logistic regression or simple decision trees. The higher explanation time is primarily due to:

- The sophistication of ensemble models such as XGBoost.
- The application of SHAP, which estimates feature contributions via permutations using game theory.
- The extra layer of interpretability afforded by counterfactual and local surrogate models.

In spite of this, the time taken for explanation is still within acceptable limits for real-time financial use, since most banking and payment systems have a few seconds' tolerance to latency per transaction. Notably, the added interpretability:

- Increases analyst confidence in model outputs
- Assists in regulatory compliance (e.g., GDPR, RBI guidelines),
- Gives transparency to end-users and internal auditors.

This sacrifice of speed for trust is warranted, particularly in high-risk, high-regulation environments such as credit card fraud detection.

Limitations

- The time for explanation, while tolerable, can grow linearly with feature complexity in extremely large datasets.
- A few descriptions might still remain too technical for layman users, even with a push for simplifying output in natural language.
- Counterfactual generation is computationally costly, particularly on high-dimensional data.
- Biases in the training data may still impact both prediction and explanation quality if not carefully addressed during data pre-processing.

References

1. Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020). Explainable AI in Fintech Risk Management. *Frontiers in Artificial Intelligence*, 3, 26. <https://doi.org/10.3389/frai.2020.00026>
2. Wu, T., & Wang, Y. (2021). Locally Interpretable One-Class Anomaly Detection for Credit Card Fraud Detection", arXiv:2108.02501
3. Z. Yang, A. Sudjianto, X. Li, and A. Zhang, "Inherently Interpretable Tree Ensemble Learning," arXiv preprint arXiv:2410.19098, Oct. 2024."
4. Ayalew et al. (2021) ," Deep autoencoders for anomaly detection in fraud and leveraged LIME to explain rare fraudulent behavior".
5. Zhang et al. (2022),"graph-based methods for modeling transaction relationships, providing new avenues for using XAI in structured data scenarios".
6. Arya et al. (2021) argued that "one explanation does not fit all" and proposed user-centric explanations depending on stakeholder roles (e.g., customers, investigators).
7. Molnar (2020), "exhaustive overview of model interpretability techniques and their relevance in domains such as finance and healthcare".
8. Verma et al. (2020),"A systematic review of counterfactual explanations,"
9. Tonekaboni et al. (2020), "Contextual explainability in high-risk domains",
10. Carcillo et al. (2020), " Unsupervised (e.g., clustering) and supervised learning (e.g., decision trees) for fraud detection,".
11. Pawelczyk et al. (2021), "Causal models for generating explanations",
12. Guidotti et al. (2021)," Methods for explaining black-box models, identifying scalability and usability as ongoing challenges in deploying XAI in real-world systems.
13. Das et al. (2023)," opportunities and challenges of using XAI in fraud detection".
14. Bhatt et al. (2020) "real-world deployments of XAI".
15. Binns et al. (2021) ' perceptions of fairness in algorithmic decisions, underlining the psychological importance of interpretable outputs in fraud systems".
16. Ahmad & Eckert (2022) "Survey on interpretable machine learning in network security, with relevance to transaction security and fraud."
17. Reddy, P.R.S et.ai. (2025) , "Adaptive and Scalable Cloud Data Sharing Framework with Quantum-Resistant Security, Decentralized Auditing, and Machine Learning-Based Threat Detection, *International Journal of Advanced Computer Science and Applications*, Volume 16, Issue 2, 2025, Pages 1042-1047