

Transparency And Privacy The Role Of Explainable AI And Federated Learning In Financial Fraud Detection

Dr.K.Raghavendar¹, K.Sai Sruthi², G.Shirisha³, K.Nikhil⁴

¹ Associate Professor, Department of Computer Science and Engineering

^{2,3,4} B.Tech Students, Department of Computer Science and Engineering

Teegala Krishna Reddy Engineering College, Telangana, India

Abstract - Financial fraud has become a major challenge for financial institutions due to the rapid growth of digital transactions and online banking systems. Traditional fraud detection techniques often rely on complex deep learning models such as Deep Neural Networks (DNN) and Recurrent Neural Networks (RNN). Although these models provide high prediction accuracy, they typically operate as black-box systems, making it difficult to understand how decisions are made. Additionally, centralized data processing in traditional systems raises significant concerns related to data privacy and security. To address these challenges, this research proposes a financial fraud detection framework that integrates Explainable Artificial Intelligence (XAI) and Federated Learning (FL). The system utilizes the PaySim dataset to analyze transaction patterns and identify fraudulent activities. Interpretable machine learning algorithms such as Decision Trees, Random Forests, Gradient Boosting Machines, and XGBoost are employed to improve model transparency and prediction accuracy. Furthermore, Federated Learning enables decentralized model training across multiple devices without sharing raw financial data, thereby preserving user privacy. Explainability techniques such as SHAP values are used to interpret model predictions and highlight feature importance in fraud detection. Experimental results demonstrate that the proposed approach improves transparency, privacy preservation, and fraud detection performance compared to traditional methods. The integration of Explainable AI and Federated Learning provides a robust and secure framework for developing trustworthy financial fraud detection systems.

Key Words: Financial Fraud Detection, Explainable Artificial Intelligence (XAI), Federated Learning (FL), Decision Tree, Random Forest, Gradient Boosting, XGBoost, Machine Learning, Privacy-Preserving Learning, SHAP Explainability.

1.INTRODUCTION

Financial fraud has emerged as a significant challenge in modern financial systems due to the rapid expansion of digital payment platforms, online banking, and mobile financial services. As digital financial transactions continue to increase globally, financial institutions process enormous volumes of transaction data every day. While these technologies improve accessibility, efficiency, and convenience, they also create new opportunities for

fraudulent activities. Fraudulent practices such as identity theft, unauthorized transactions, and payment manipulation can lead to substantial financial losses for both individuals and organizations.

Traditionally, fraud detection systems relied on rule-based mechanisms and manual verification techniques. Although these approaches were effective for small-scale transaction monitoring, they struggle to handle large-scale financial datasets and evolving fraud patterns. In recent years, machine learning and artificial intelligence techniques have been increasingly adopted to automate fraud detection and improve prediction accuracy. Models such as Deep Neural Networks (DNN) and Recurrent Neural Networks (RNN) have shown strong performance in identifying complex fraud patterns. However, these deep learning models often function as “black-box” systems, making it difficult to interpret their decision-making processes. Additionally, many existing fraud detection frameworks rely on centralized data storage, which raises serious concerns regarding data privacy and security. To address these challenges, this research proposes an advanced fraud detection framework that integrates Explainable Artificial Intelligence (XAI) and Federated Learning (FL). Explainable AI improves transparency by providing interpretable insights into model predictions, while Federated Learning enables decentralized model training without transferring sensitive financial data to a central server. By combining these two approaches, the proposed system aims to enhance both the transparency and privacy of fraud detection systems while maintaining high predictive performance.

1.1 Motivation

The increasing volume of digital financial transactions has significantly increased the risk of fraudulent activities. Financial institutions must analyze massive datasets to identify suspicious transaction patterns and prevent fraud effectively. However, many traditional machine learning models lack transparency, making it difficult for analysts to understand how predictions are generated. This lack of interpretability reduces trust in automated fraud detection systems and creates challenges for regulatory compliance.

Furthermore, financial transaction data contains highly sensitive information, including personal and financial details of users. Centralized data processing exposes such

data to potential security breaches and privacy risks. Therefore, there is a strong need for fraud detection systems that not only provide accurate predictions but also ensure data privacy and model interpretability. The motivation behind this research is to develop a transparent and privacy-preserving fraud detection framework by integrating Explainable AI techniques with Federated Learning.

1.2 Problem Statement

Despite the progress made in machine learning-based fraud detection, several challenges remain unresolved. Deep learning models such as DNN and RNN can achieve high accuracy but often lack interpretability, making it difficult for financial institutions to understand and justify automated decisions. Moreover, many existing fraud detection systems rely on centralized data collection for model training. Storing large volumes of sensitive financial data in centralized systems increases the risk of data breaches and privacy violations. Therefore, there is a need for an advanced fraud detection system that not only achieves high detection accuracy but also ensures transparency, interpretability, and strong data privacy protection.

1.3 Objectives of the Study

The primary objective of this research is to develop a financial fraud detection system that integrates Explainable Artificial Intelligence and Federated Learning to enhance transparency, interpretability, and data privacy. The study aims to develop machine learning models capable of accurately detecting fraudulent financial transactions using the PaySim dataset. It also focuses on implementing interpretable algorithms such as Decision Trees, Random Forests, Gradient Boosting, and XGBoost to improve model explainability. Additionally, Federated Learning is applied to enable decentralized model training without sharing sensitive financial data. By achieving these objectives, the proposed system aims to provide a reliable, efficient, and privacy-preserving fraud detection framework suitable for real-world financial applications.

2. PROPOSED SYSTEM

The proposed system introduces a transparent and privacy-preserving framework for financial fraud detection by integrating Explainable Artificial Intelligence (XAI) and Federated Learning (FL) with interpretable machine learning algorithms. The system is designed to detect fraudulent financial transactions while ensuring that the decision-making process is understandable and that sensitive financial data remains protected.

Traditional fraud detection systems rely heavily on deep learning models such as Deep Neural Networks (DNN) and Recurrent Neural Networks (RNN). Although these models

can achieve high accuracy, they often lack interpretability and require centralized data collection for model training. This raises concerns related to data privacy, regulatory compliance, and trust in automated decision systems. To address these challenges, the proposed approach utilizes interpretable machine learning algorithms including Decision Trees, Random Forests, Gradient Boosting Machines, and XGBoost. These algorithms provide better transparency in model decisions while maintaining strong predictive performance. Additionally, the integration of Federated Learning allows the model to be trained across decentralized devices without transferring raw transaction data to a central server. Explainable AI techniques such as SHAP (Shapley Additive Explanations) are used to interpret model predictions by identifying the importance of different transaction features. This helps financial analysts understand the reasoning behind fraud detection decisions.

The system processes transaction data from the PaySim dataset, which simulates real-world mobile financial transactions. The dataset undergoes preprocessing, feature extraction, and model training before generating fraud predictions.

2.1 Data Collection and Dataset Description

The proposed system utilizes the PaySim dataset, which is a synthetic financial dataset generated from real mobile money transaction logs. The dataset contains several transaction attributes such as transaction type, amount, old balance, new balance, and transaction destination. This dataset provides a realistic representation of financial transaction patterns, enabling the system to learn and identify suspicious behaviors associated with fraudulent activities. The dataset is divided into training and testing sets to evaluate the performance of different machine learning models.

2.2 Data Preprocessing

Data preprocessing is an essential step in the fraud detection process. Raw financial transaction data may contain missing values, redundant attributes, or inconsistent formats that can negatively impact model performance. During preprocessing, the system performs several operations including data cleaning, feature selection, normalization, and removal of irrelevant attributes. These steps ensure that the dataset is structured and suitable for machine learning model training. The preprocessed data is then transformed into a format compatible with machine learning algorithms to improve prediction accuracy and computational efficiency.

2.3 Machine Learning Models for Fraud Detection

The proposed system employs multiple machine learning algorithms to detect fraudulent transactions. These algorithms are selected based on their interpretability, performance, and ability to handle complex datasets.

Decision Trees provide a hierarchical structure for decision-making and allow easy interpretation of classification rules. Random Forest improves prediction accuracy by combining multiple decision trees and reducing overfitting. Gradient Boosting Machines enhance model performance by sequentially combining weak learners into a strong predictive model. XGBoost further improves gradient boosting by introducing optimization techniques and parallel computation. These models analyze transaction patterns and classify each transaction as either fraudulent or legitimate.

2.4 Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence is incorporated into the proposed system to improve transparency in model predictions. Many machine learning models operate as black boxes, making it difficult for users to understand the reasoning behind predictions. To overcome this limitation, the system uses SHAP (Shapley Additive Explanations) to interpret model outputs. SHAP values measure the contribution of each feature toward the final prediction. By analyzing SHAP values, financial analysts can identify which transaction attributes play the most important role in determining whether a transaction is fraudulent. This improves trust in automated fraud detection systems and supports regulatory compliance.

2.5 Federated Learning for Privacy Preservation

Federated Learning is integrated into the proposed system to protect sensitive financial data during model training. Instead of transferring raw data to a central server, the training process occurs locally on multiple devices or nodes.

Each local device trains a model using its own data and sends only model updates (such as gradients or parameters) to the central server. The server aggregates these updates to create a global model that benefits from the collective knowledge of all devices. This approach ensures that sensitive financial data never leaves its original location, thereby preserving user privacy and reducing the risk of data breaches.

2.6 System Architecture

The architecture of the proposed financial fraud detection system consists of several interconnected components including data collection, preprocessing, machine learning model training, explainability module, and federated learning framework. The system first collects transaction data from the PaySim dataset. The data then undergoes preprocessing to remove inconsistencies and prepare it for

model training. Multiple machine learning models are trained using the processed dataset to classify transactions as fraudulent or legitimate.

The Explainable AI module generates SHAP values to interpret model predictions and identify feature importance. Simultaneously, the Federated Learning framework enables decentralized model training while preserving data privacy. Finally, the system outputs fraud detection results and provides explanations for each prediction.

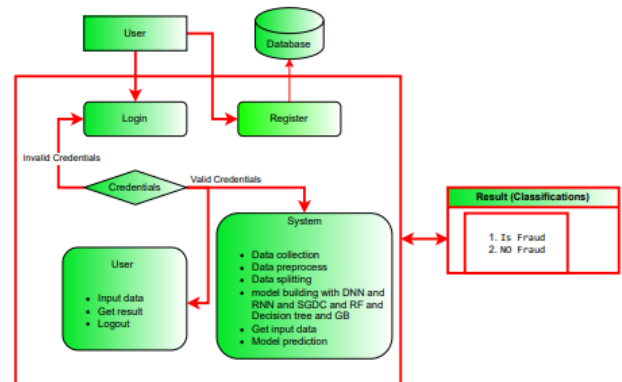


Fig. 1: System Architecture of the Proposed Financial Fraud Detection System

3. IMPLEMENTATION DETAILS

The implementation of the proposed financial fraud detection system focuses on integrating machine learning techniques with Explainable Artificial Intelligence (XAI) and Federated Learning (FL) to accurately detect fraudulent financial transactions while ensuring data privacy. The system is developed using the Python programming language along with several machine learning libraries such as Scikit-learn, TensorFlow, and XGBoost. The PaySim dataset is used for training and evaluating the predictive models. The implementation process involves several stages including dataset collection, data preprocessing, model training, explainability analysis, and prediction generation through a web-based interface. Each stage contributes to improving the efficiency, reliability, and transparency of the fraud detection system.

3.1 System Modules

The proposed system consists of multiple functional modules that collectively perform fraud detection tasks. These modules ensure proper data processing, model training, prediction generation, and result interpretation.

The User Interface Module allows users to interact with the system by providing functionalities such as user registration, login, dataset upload, model selection, and transaction input for prediction. The Dataset Upload and Management Module enables users to upload the PaySim dataset. This module

validates the dataset format and allows users to view transaction records for analysis. The Data Preprocessing Module prepares the dataset for machine learning model training. It performs operations such as removing irrelevant attributes, handling missing values, encoding categorical variables, and feature extraction to improve model accuracy. The Machine Learning Training Module trains multiple machine learning algorithms using the processed dataset to classify transactions as fraudulent or legitimate. The Prediction Module allows users to input transaction details and obtain fraud prediction results. Finally, the Explainability and Visualization Module interprets model predictions using SHAP values and presents performance metrics such as accuracy scores, confusion matrices, and ROC curves.

3.2 Dataset Processing

The PaySim dataset is used for training and evaluating the fraud detection models. This dataset simulates mobile financial transactions and contains features such as transaction type, transaction amount, sender balance, receiver balance, and transaction destination.

Before model training, the dataset undergoes several preprocessing steps to ensure data quality. These steps include removing redundant attributes, converting categorical variables into numerical form, and normalizing numerical features. The processed dataset is then divided into training and testing subsets to evaluate model performance.

3.3 Machine Learning Algorithms

Several machine learning algorithms are implemented in the proposed system to detect fraudulent transactions. These algorithms are selected based on their ability to handle large datasets and provide reliable predictions. A Deep Neural Network (DNN) is used to capture complex patterns in transaction data using multiple hidden layers. A Recurrent Neural Network (RNN) is implemented to analyze sequential relationships between transactions, which is useful for identifying fraud patterns over time. The Stochastic Gradient Descent Classifier (SGDC) is employed as an efficient optimization-based classification model that processes data iteratively, making it suitable for large-scale datasets. The Decision Tree algorithm is used as an interpretable model that classifies transactions based on hierarchical decision rules.

The Random Forest algorithm improves prediction performance by combining multiple decision trees, thereby reducing overfitting and improving classification accuracy. Additionally, Gradient Boosting and XGBoost algorithms are implemented to enhance predictive performance by sequentially combining weak learners to minimize classification errors.

3.4 Explainable AI Implementation

To improve transparency in fraud detection decisions, the system integrates Explainable Artificial Intelligence techniques using SHAP (Shapley Additive Explanations). SHAP values quantify the contribution of each feature toward the final prediction generated by the machine learning model. SHAP analysis helps visualize feature importance and explains how different transaction attributes influence fraud predictions. For instance, features such as transaction amount, balance difference, and transaction type may significantly impact the classification outcome. These insights enable financial analysts to better understand the model's decision-making process.

3.5 Federated Learning Implementation

Federated Learning is incorporated into the proposed system to ensure privacy-preserving model training. Instead of transferring raw financial transaction data to a centralized server, the training process occurs locally on multiple devices or nodes. Each local model is trained using the available data and only the model updates are transmitted to a central server. The server aggregates these updates to create a global model that benefits from knowledge learned across all devices. This decentralized learning approach ensures that sensitive financial data remains on local devices, thereby reducing the risk of data breaches.

3.6 System Interface and Output

The system is deployed through a web-based interface developed using Python frameworks. The interface allows users to upload datasets, select machine learning algorithms, and perform fraud prediction tasks. The system generates prediction outputs indicating whether a transaction is fraudulent or legitimate. In addition, the interface provides visual outputs such as confusion matrices, ROC curves, and SHAP feature importance graphs to help users interpret model performance.

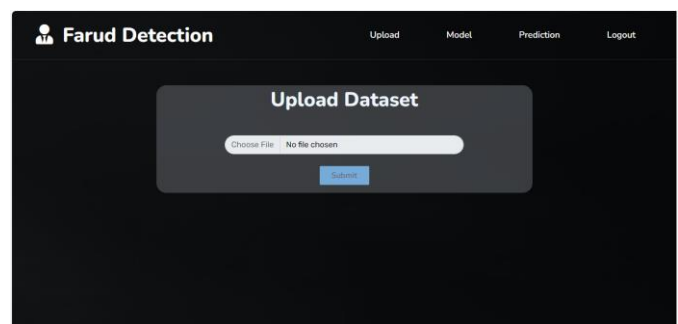


Fig. 2: Dataset Upload and Model Selection Interface

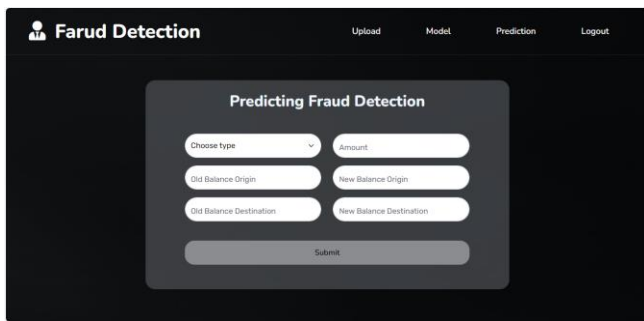


Fig. 3: Fraud Prediction Result Interface

4. RESULTS AND PERFORMANCE ANALYSIS

The performance of the proposed financial fraud detection system was evaluated using the PaySim dataset. Multiple machine learning algorithms were implemented and compared to analyze their effectiveness in detecting fraudulent transactions. The evaluation was conducted using standard performance metrics including accuracy, precision, recall, F1-score, and confusion matrix. The dataset was divided into training and testing subsets to ensure unbiased model evaluation. After preprocessing, several machine learning algorithms such as Deep Neural Network (DNN), Recurrent Neural Network (RNN), Stochastic Gradient Descent Classifier (SGDC), Decision Tree, Random Forest, Gradient Boosting, and XGBoost were trained and tested for fraud classification.

4.1 Performance Metrics

To evaluate the effectiveness of the fraud detection models, several performance metrics were used. Accuracy measures the proportion of correctly classified transactions. Precision indicates the percentage of predicted fraudulent transactions that are actually fraudulent. Recall measures the ability of the model to correctly identify fraudulent transactions. The F1-score provides a balanced measure of precision and recall, which is particularly useful for handling imbalanced datasets. Additionally, the confusion matrix was used to analyze the classification results by identifying true positives, true negatives, false positives, and false negatives.

4.2 Model Performance Evaluation

The experimental results demonstrate that machine learning models are capable of effectively detecting fraudulent transactions. Among the evaluated algorithms, ensemble learning techniques such as Random Forest, Gradient Boosting, and XGBoost achieved higher prediction accuracy compared to traditional models. Decision Tree models provided interpretable classification rules, while Random Forest improved model stability by combining multiple decision trees and reducing overfitting. Gradient Boosting and XGBoost further enhanced prediction performance by sequentially minimizing classification errors. Although deep

learning models such as DNN and RNN also achieved good detection performance, they lacked interpretability when compared with tree-based ensemble models.

4.3 Explainability Analysis

Explainable Artificial Intelligence techniques were applied to interpret the predictions generated by the machine learning models. SHAP (Shapley Additive Explanations) values were used to measure the contribution of different features in determining fraud predictions. The analysis revealed that features such as transaction amount, balance difference, and transaction type significantly influence the detection of fraudulent transactions. SHAP-based explanations improve transparency by providing insights into how the model arrives at its predictions.

4.4 Comparative Analysis

A comparative analysis of all implemented models indicates that ensemble learning algorithms outperform individual models in terms of prediction accuracy and stability. Random Forest and XGBoost achieved the best performance among the evaluated algorithms. The integration of Explainable AI improves model transparency, while Federated Learning ensures privacy-preserving model training. Overall, the proposed system demonstrates effective fraud detection capability while maintaining transparency and data privacy, making it suitable for real-world financial security applications.

5. CONCLUSIONS

Financial fraud detection has become increasingly critical with the rapid growth of digital financial transactions and online banking systems. Traditional fraud detection methods often rely on complex deep learning models that achieve high accuracy but lack transparency and depend on centralized data processing, which raises concerns regarding interpretability and data privacy.

In this research, a financial fraud detection framework integrating Explainable Artificial Intelligence (XAI) and Federated Learning (FL) was proposed to address these challenges. The system utilizes the PaySim dataset to analyze transaction patterns and detect fraudulent activities using multiple machine learning algorithms, including Decision Tree, Random Forest, Gradient Boosting, XGBoost, Deep Neural Network (DNN), Recurrent Neural Network (RNN), and Stochastic Gradient Descent Classifier (SGDC).

The experimental results show that ensemble learning methods such as Random Forest and Gradient Boosting achieve higher prediction accuracy and stability compared to traditional models. Additionally, the use of Explainable AI techniques such as SHAP values improves the transparency of model predictions by identifying the contribution of

different transaction features. This enables financial analysts to better understand the model's decision-making process.

Furthermore, the integration of Federated Learning enables privacy-preserving model training by allowing decentralized learning across multiple devices without sharing raw financial data. This approach significantly reduces the risk of data breaches while maintaining effective fraud detection performance.

6. FUTURE WORK

Although the proposed financial fraud detection system improves transparency, accuracy, and privacy preservation, there are several opportunities for further enhancement. Future research can focus on integrating more advanced machine learning and deep learning techniques to improve the detection of complex fraud patterns. Hybrid models that combine interpretable algorithms with advanced architectures such as transformer-based models or graph neural networks could be explored to capture more complex relationships within financial transaction data. Another potential direction is the development of real-time fraud detection systems capable of analyzing streaming transaction data and identifying suspicious activities instantly. Such systems would enable financial institutions to prevent fraudulent transactions before they are completed, thereby reducing financial losses and enhancing system security. In addition, the federated learning framework used in this research can be further strengthened by incorporating techniques such as differential privacy and secure aggregation. These approaches can provide additional protection for sensitive financial data while maintaining collaborative model training across decentralized devices. Future research may also involve evaluating the proposed system on large-scale real-world financial datasets from multiple institutions to assess its scalability and robustness. Moreover, integrating advanced visualization and explainability tools could further improve the interpretability of model predictions for financial analysts and regulatory authorities.

REFERENCES

- [1] UK Finance, "Annual Fraud Report 2022," UK Finance, 2022. [Online]. Available: <https://www.ukfinance.org.uk/policy-and-guidance/reports-and-publications/annual-fraud-report-2022>
- [2] A. Abdallah, M. A. Maarof, and A. Zainal, "Fraud Detection System: A Survey," *Journal of Network and Computer Applications*, vol. 68, pp. 90–113, Jun. 2016.
- [3] A. Pascual, K. Marchini, and S. Miller, "Identity Fraud: Securing the Connected Life," *Javelin Strategy & Research*, 2017.
- [4] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data Mining for Credit Card Fraud: A Comparative Study," *Decision Support Systems*, vol. 50, no. 3, pp. 602–613, Feb. 2011.
- [5] L. T. Rajesh, T. Das, R. M. Shukla, and S. Sengupta, "Give and Take: Federated Transfer Learning for Industrial IoT Network Intrusion Detection," *arXiv preprint arXiv:2310.07354*, 2023.
- [6] S. Vyas, A. N. Patra, and R. M. Shukla, "Histopathological Image Classification and Vulnerability Analysis Using Federated Learning," *arXiv preprint arXiv:2306.05980*, 2023.
- [7] R. J. Bolton and D. J. Hand, "Statistical Fraud Detection: A Review," *Statistical Science*, vol. 17, no. 3, pp. 235–255, Aug. 2002.
- [8] H. van Driel, "Financial Fraud, Scandals, and Regulation: A Conceptual Framework and Literature Review," *Business History*, vol. 61, no. 8, pp. 1259–1299, Nov. 2019.
- [9] G. M. Trompeter, T. D. Carpenter, N. Desai, K. L. Jones, and R. A. Riley, "A Synthesis of Fraud-Related Research," *Auditing: A Journal of Practice & Theory*, vol. 32, no. S1, pp. 287–321, May 2013.
- [10] P. Raghavan and N. E. Gayar, "Fraud Detection Using Machine Learning and Deep Learning," in *Proceedings of the International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, 2019, pp. 334–339.