

# Implementing Ethical AI Governance Framework with SHAP and LIME for Enhanced Cybersecurity in Financial Institutions

Amrutha B

Krupanidhi Degree College

\*\*\*

**Abstract** - Financial institutions' capacity to identify and address cyberthreats has been greatly improved by the incorporation of artificial intelligence (AI). The volume and complexity of financial data have increased due to the growth of digital transactions, making sophisticated cybersecurity measures necessary. However, serious questions about accountability, transparency, and ethical governance have been raised by the opaque nature of AI-driven decision-making, known as the "black box" problem. To tackle these issues, this research uses the Synthetic Financial Datasets for Fraud Detection to develop an ethical AI governance framework that integrates explainable AI (XAI) methods, namely SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), into a fraud detection system. A high-dimensional, unbalanced financial dataset was used to train and assess the XGBoost classifier. According to experimental results, combining SHAP and LIME greatly improved the model's interpretability while maintaining a high accuracy rate of over 99.97% with strong precision and recall. The framework promotes stakeholder trust, fair model validation, and bias detection.

**Key Words:** AI in cybersecurity, explainable AI, SHAP, LIME, ethical AI governance, fraud detection

## 1. INTRODUCTION

Financial institutions' operations have been completely transformed by the incorporation of Artificial Intelligence (AI), which offers strong security measures, increased efficiency, and better customer experiences. However, the use of AI systems raises issues with accountability, transparency, and ethics, particularly in the field of cybersecurity. Using explainable AI (XAI) methods like SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) in an ethical AI governance framework is essential to resolving these issues. In order to ensure that financial institutions can effectively manage and trust their AI-driven cybersecurity systems, this framework attempts to clarify AI decision-making processes.

AI is being used by financial institutions more and more to counteract sophisticated cyberthreats. Cybersecurity systems powered by AI are able to analyse enormous volumes of data, spot irregularities, and react to threats instantly. But because of their intricacy, these AI models frequently become "black boxes," making it difficult for stakeholders to comprehend how they make decisions. This opacity can result in ethical quandaries, regulatory non-compliance, and mistrust, especially when AI decisions affect stakeholders and customers.

By offering insights into AI model predictions, explainable AI (XAI) techniques like SHAP and LIME have surfaced to address these issues. LIME approximates the model locally to comprehend its behaviour around a specific instance, whereas SHAP assigns an importance value to each feature, indicating its contribution to a specific prediction. Financial institutions can improve stakeholder understanding and trust by incorporating these strategies into their AI systems to increase transparency.

### 1.1 Background

Traditional rule-based security systems are frequently unable to identify new attack vectors in real-time due to the increasing sophistication of cyberattacks. Systems with AI capabilities have become effective instruments to counter these dangers. These models, however, frequently operate as "black boxes," offering little to no justification for their predictions. This is especially problematic in the financial industry, where decision traceability is crucial.

### 1.2 Problem Statement

AI models such as XGBoost have poor interpretability even though they provide better predictive performance in identifying fraudulent financial transactions. This opacity can create ethical dilemmas, impair regulatory compliance,

and breed mistrust among stakeholders. A thorough framework that uses explainable AI techniques to guarantee high performance while also encouraging accountability and transparency is desperately needed.

### 1.3 Research Objectives

The primary objectives of this research are: (i) Select a suitable financial cybersecurity dataset for training AI models; (ii) Train machine learning models (XGBoost) to identify suspicious transactions; (iii) Use SHAP and LIME to make AI models interpretable; (iv) Perform model fairness and bias analysis; and (v) Define metrics to measure model performance based on accuracy, interpretability, and ethical implications.

## 2. LITERATURE REVIEW

### 2.1 Explainable AI in Financial Fraud Detection

The problem of irregularities in financial transactions and the opaqueness of AI-based machine learning (ML) detection models are examined in prior work. Sabharwal et al. [1] emphasise how crucial XAI methods—in particular, SHapley Additive exPlanations (SHAP)—are for giving financial managers clear justifications for anomaly detections produced by AI, extending its application through action design research methodology. Mohsin and Nasim [3] conducted a systematic review of XAI methods in financial fields like fraud detection and credit scoring, finding that post-hoc interpretability techniques are heavily used, with SHAP being the most prevalent method, followed by attention mechanisms and feature importance analysis.

Yang et al. [6] provide a comprehensive survey on explainable AI approaches, classifying methods into three distinct taxonomies and discussing the role of techniques like SHAP and LIME in providing quantitative explanations, improving interpretability, and lowering risks in financial applications. Gayathri et al. [22] demonstrated that a hybrid explainable AI model combining SHAP with Transformer architectures achieved an accuracy of 97.71% with an F1-score of 0.96, with fairness assessments revealing Demographic Parity of 96.40% and Equal Opportunity of 94.40%.

### 2.2 Ethical AI Governance Frameworks

The ESG-AI framework proposed by Lee et al. [2] offers an organised method for assessing the material impact of AI and a company's dedication to responsible AI, based on insights from interactions with 28 businesses, made publicly available in April 2024. Financial services have spearheaded efforts to govern AI since the early 2010s; Sayles [4] highlights that organisations are increasingly assembling cross-functional teams comprising legal specialists, data scientists, and AI ethicists to address explainability, bias prevention, and risk mitigation. Gasser and Almeida [11] proposed a foundational layered model for AI governance addressing information asymmetries between developers, consumers, and policymakers.

Kandeel et al. [12] provide a broad coverage of AI governance issues analysed through international initiatives including the European AI Act, UNESCO AI Ethics Recommendation, and OECD AI Principles, identifying key governance principles such as transparency, accountability, data privacy, and inclusivity. Rajagopal et al. [14] propose a smart governance framework for AI in public administration, while Cooreman and Zhu [15] offer alternative regulation approaches by critically examining existing AI ethics frameworks. Deshpande [25] specifically addresses regulatory compliance challenges of AI in finance, proposing a novel framework with mathematical models to forecast regulatory alignment and potential breaches.

### 2.3 AI Security and Cybersecurity Applications

Mhammad et al. [8] investigate how Large Language Models (LLMs) can improve compliance in financial transactional platforms in the area of money laundering detection through differential governance, combining anomaly detection and generative AI. PenTest++ by S. and Mitchell [21] demonstrates how AI-driven systems can supplement human expertise in cybersecurity through automation of penetration testing tasks including reconnaissance, scanning, and exploitation. Al-Sinani et al. [30] examine the potential of generative AI to improve cybersecurity through ethical hacking, critically evaluating risks such as data biases, over-reliance on technology, and misuse of AI.

**Table 1: Summary of Critical Analysis of Existing Studies**

| Reference               | Contributions                                      | Methodology                                 | Limitations                       | Advantages                                     |
|-------------------------|----------------------------------------------------|---------------------------------------------|-----------------------------------|------------------------------------------------|
| Sabharwal et al. (2024) | Extends XAI for managers in finance                | Case studies, XAI models                    | Limited to managerial perspective | Practical insights for financial professionals |
| Lee et al. (2024)       | Proposes responsible AI assessment integrating ESG | Framework development, qualitative analysis | Lacks empirical validation        | Strong theoretical foundation                  |
| Mohsin & Nasim (2025)   | Systematic review of explainable AI in finance     | Literature review                           | Lacks experimental verification   | Comprehensive XAI trends analysis              |
| Sayles (2024)           | Overview of AI governance and model risk           | Theoretical analysis                        | Limited empirical studies         | Covers broad governance perspective            |
| Saleem et al. (2025)    | Responsible AI challenges in fintech               | Case studies, industry reports              | Lacks technical details           | Relevant to financial technology firms         |
| Yang et al. (2023)      | Survey on XAI approaches and limitations           | Systematic literature review                | Theoretical study                 | Broad understanding of XAI                     |

## 2.4 Research Gap

Despite the growing body of literature on explainable AI and ethical governance frameworks, several critical gaps remain unaddressed. First, most existing studies treat XAI and ethical governance as separate concerns—works focusing on SHAP and LIME interpretability (e.g., [1], [3], [6]) do not integrate these into a comprehensive governance framework with regulatory compliance checks, while governance-focused works (e.g., [4], [11], [12]) lack technical XAI implementation. Second, the majority of fraud detection studies using XGBoost or ensemble methods do not simultaneously address explainability, fairness/bias detection, and regulatory alignment (GDPR, AI Act) within a single unified pipeline.

Third, existing hybrid XAI models [22] are tested on general datasets without specific focus on highly imbalanced financial cybersecurity datasets where class ratios can be as extreme as 0.13% fraud vs 99.87% non-fraud. Fourth, while studies acknowledge the computational overhead of SHAP and LIME in real-time systems [6], no prior work has empirically quantified the accuracy-interpretability trade-off in the specific context of financial fraud detection using a governance-compliant architecture. This research directly addresses these gaps by: (i) integrating both SHAP and LIME into a unified ethical AI governance pipeline; (ii) evaluating the framework on a severely imbalanced financial fraud dataset; (iii) quantifying the performance trade-off between the original and XAI-filtered models; and (iv) systematically mapping the framework outputs to GDPR and regulatory compliance requirements.

## 3. METHODOLOGY

Data collection, model development, explainability integration, evaluation, and validation are the various stages of the methodology for implementing an Ethical AI Governance Framework with SHAP and LIME for Enhanced Cybersecurity in Financial Institutions. This methodical approach guarantees that AI models utilised in financial cybersecurity are both ethically compliant with AI governance principles and comprehensible.

### 3.1 Data Collection and Preprocessing

The Financial Fraud Detection Dataset from Kaggle was selected, which contains transactional data with fraud labels. Preprocessing steps included: (i) Data cleaning to handle missing values, outliers, and inconsistent entries; (ii) Feature engineering to extract important features such as transaction type, amount, time of transaction, and device information; (iii) Normalisation using Standard Scaler to improve model performance; (iv) Fraud label encoding to transform categorical labels (fraud/non-fraud) into numerical values; and (v) Train-test split of 70% training and 30% testing.

### 3.2 ML/AI Model Development

Based on its high predictive power, efficiency, and resilience, Extreme Gradient Boosting (XGBoost) was chosen as the main classification model. XGBoost creates decision trees in a sequential fashion, with each new tree aiming to fix the mistakes of the ones before it. Among XGBoost's benefits are: using regularisation to avoid overfitting, naturally handling missing values, fast speed and scalability, and better results on tabular and structured data. Hyperparameter tuning was performed using grid search with five-fold cross-validation.

Two cutting-edge explainability techniques were employed to guarantee transparency and interpretability. SHAP (SHapley Additive exPlanations) is a game-theoretic method that gives each feature an importance value by considering every possible feature combination. Global feature importance calculations and visualisation of individual feature contributions were performed using SHAP. LIME (Local Interpretable Model-Agnostic Explanations) focuses on local interpretability by approximating the model locally using a simple, interpretable surrogate model to explain individual predictions.

### 3.3 Evaluation Metrics

To evaluate the proposed system, the following metrics were used: Accuracy ( $TP+TN / TP+TN+FP+FN$ ) for overall correctness; Sensitivity/Recall ( $TP / TP+FN$ ) for the ability to detect fraudulent transactions; Precision ( $TP / TP+FP$ ) to minimise false positives; and F1-Score (harmonic mean of precision and recall) for a balanced metric that handles class imbalance. Additionally, confusion matrices and ROC-AUC curves were used for visualisation and comparative analysis.

## 4. EXPERIMENTAL RESULTS

### 4.1 Experimental Setup

All model training and evaluation procedures were conducted on Google Colab with: Intel Xeon CPU at 2.30GHz, NVIDIA Tesla T4 GPU, 12 GB RAM, Linux-based virtual environment, Python 3.9. Libraries used included Scikit-learn, XGBoost, SHAP, LIME, Seaborn, Matplotlib, NumPy, and Pandas.

### 4.2 Dataset Description

The Synthetic Financial Dataset for Fraud Detection (SFD-FFD) served as the main dataset. The dataset contains 1,908,786 total transactions, of which 2,435 (~0.13%) were fraudulent and 1,906,351 (~99.87%) were non-fraudulent. Data preprocessing included: data cleaning to eliminate outliers, duplicates, and null entries; feature encoding using one-hot and label encoding; feature scaling using standard scaling; and class imbalance handling through class-weight tuning. The dataset was split into 80% training and 20% testing.

### 4.3 Results

The XGBoost model was selected due to its proven efficiency in handling large datasets with class imbalance and its compatibility with explainability frameworks like SHAP and LIME. The training process used a grid search approach for hyperparameter tuning, optimising for maximum recall and F1-score. Five-fold cross-validation was used during model training to guarantee consistency across data splits.

The SHAP Summary Plot identified features like "transaction amount," "time of day," and "payment method" as important fraud indicators by visualising feature contributions across all predictions. LIME provided localised interpretations for specific fraudulent transactions, identifying feature thresholds and trends that contribute to fraud labels.

**Table 2: Final Performance Comparison**

| Model Version       | Accuracy | Precision | Recall   | F1-Score |
|---------------------|----------|-----------|----------|----------|
| Original Model      | 0.999754 | 0.958469  | 0.843532 | 0.897335 |
| SHAP-Filtered Model | 0.999751 | 0.955390  | 0.844353 | 0.896446 |
| LIME-Filtered Model | 0.999751 | 0.955390  | 0.844353 | 0.896446 |

The precision, recall, and F1-score metrics show the true performance insights, even though the accuracy is constantly high across all models because of class imbalance. The modest performance variations following the application of SHAP and LIME demonstrate the trade-off between transparency and a slight decline in predictive power. Nonetheless, the incorporation of explainability techniques is justified by the improvements in model interpretability.

#### 4.4 Comparison with Baseline Methods

Traditional models such as Logistic Regression and Random Forest were implemented as baselines for performance benchmarking. Logistic regression frequently fails to capture the intricate, non-linear relationships found in high-dimensional fraud datasets. Random Forests may not be as successful as gradient boosting techniques in optimising for highly skewed data distributions. The XGBoost model employed in this investigation performed better on important evaluation metrics like precision, recall, and F1-score, especially when enhanced with SHAP and LIME explainability tools.

## 5. CONCLUSION AND FUTURE WORK

This research presented an explainable AI-based fraud detection framework using the XGBoost classifier, supplemented with SHAP and LIME for interpretability. The framework was created and evaluated using the Synthetic Financial Datasets for Fraud Detection. Experimental results demonstrated very high accuracy (99.97%) with strong precision (0.95+), recall (0.84+), and F1-score (0.89+) values across various configurations.

The "black box" nature of XGBoost was resolved by incorporating SHAP and LIME into the AI pipeline. SHAP offered a global perspective of feature importance, while LIME provided local explanations by dissecting individual transaction predictions. This degree of explainability is essential for regulatory compliance in the financial sector, building stakeholder trust, error analysis and system auditing, and iterative model enhancements.

Future work should include: testing on real-world anonymised financial datasets; exploring ensemble approaches with LightGBM and CatBoost; implementing online learning algorithms for concept drift; adding domain-specific features such as geolocation inconsistency and device fingerprinting; deep learning sequence models for temporal fraud patterns; real-time deployment on cloud platforms; and adversarial robustness testing. The ethical implications with respect to GDPR, FATF, and the EU AI Act compliance should also be investigated.

## ACKNOWLEDGEMENT

The author expresses sincere gratitude to the Department of Computer Science and Engineering for providing the resources, computational infrastructure, and academic environment that made this research possible. Special thanks are due to the faculty members and reviewers whose valuable feedback and guidance significantly shaped the direction of this work. The author also acknowledges the open-source communities behind XGBoost, SHAP, and LIME libraries, as well as Kaggle for providing the Synthetic Financial Dataset for Fraud Detection used in the experimental evaluation.

## REFERENCES

- [1] Sabharwal, R., Miah, S.J., Wamba, S.F., et al. (2024) 'Extending application of explainable artificial intelligence for managers in financial organizations', *Annals of Operations Research*. <https://doi.org/10.1007/s10479-024-05825-9>.
- [2] Lee, S.U., Perera, H., Liu, Y., et al. (2024) 'Integrating ESG and AI: A comprehensive responsible AI assessment framework', *arXiv:2408.00965v2*.
- [3] Mohsin, M.T. and Nasim, N.B. (2025) 'Explaining the unexplainable: A systematic review of explainable AI in finance', *arXiv:2503.05966*.
- [4] Sayles, J. (2024) 'The current state of AI governance and model risk management', in *Principles of AI Governance and Model Risk Management*. Apress. <https://doi.org/10.1007/979-8-8688-09>.
- [5] Saleem, M., Chawla, C., et al. (2025) 'Responsible AI in Fintech: Addressing challenges and strategic solutions', in *Generative AI in FinTech*. Springer. [https://doi.org/10.1007/978-3-031-76957-3\\_4](https://doi.org/10.1007/978-3-031-76957-3_4).
- [6] Yang, W., Wei, Y., Wei, H. et al. (2023) 'Survey on Explainable AI: From Approaches, Limitations and Applications Aspects', *Hum-Cent Intell Syst* 3, 161-188.
- [7] Elmaghrabi, M. (2024) 'ESG risk management, sustainability-oriented firms and financial distress in the financial industry', *25th ACIT*, pp. 1-3.
- [8] Mhammad, A.F., et al. (2023) 'Generative & responsible AI - LLMs use in differential governance', *CSCI 2023*, pp. 291-295.
- [9] Agbese, M., et al. (2021) 'Governance of ethical and trustworthy AI systems', *IEEE REW 2021*, pp. 224-229.
- [10] Findlay, M. & Seah, J. (2020) 'An ecosystem approach to ethical AI and data use', *IEEE AI4G*, pp. 192-197.
- [11] Gasser, U. and Almeida, V.A.F. (2017) 'A layered model for AI governance', *IEEE Internet Computing*, 21(6), pp. 58-62.
- [12] Kandeel, M. E., et al. (2024) 'AI Governance: A General Perspective', *GCET-2024*, pp. 195-201.
- [13] Sirbu, C. L. & Mercioni, M. A. (2024) 'Fostering Trust in AI-Driven Healthcare', *ISETC 2024*, pp. 1-4.
- [14] Rajagopal, M., et al. (2023) 'A Conceptual Framework for AI Governance in Public Administration', *I-SMAC 2023*, pp. 488-495.
- [15] Cooreman, H. & Zhu, Q. (2022) 'Critical reflections on the ethical regulation of AI', *IEEE ISTAS 2022*, pp. 1-5.
- [16] McStay, A., et al. (2024) 'Ethics and empathy-based human-AI partnering', pp. 1-28.
- [17] Kim, D., Zhu, Q. & Eldardiry, H. (2023) 'Exploring approaches to AI governance: From ethics to policy', *IEEE ETHICS 2023*, pp. 1-5.
- [18] Arcilla, A. O., et al. (2023) 'Ethics in AI governance: Comparative analysis and policy recommendations for the Philippines', *ICSEC 2023*, pp. 319-326.
- [19] Ayub, Z. & Banday, M. T. (2023) 'Ethics in Artificial Intelligence: An Analysis of Ethical Issues', *STCR 2023*, pp. 1-6.
- [20] Amendi, R., et al. (2024) 'Exploring ethical implications of generative AI chatbot', *ICTIIA 2024*, pp. 1-6.
- [21] S., H., & Mitchell, C. J. (2025) 'PenTest++: Elevating Ethical Hacking with AI and Automation', *arXiv:2502.09484*.
- [22] Gayathri, J., et al. (2025) 'Ethical AI Redefined: Pioneering Transparency with Hybrid Explainable Models', *ICMSCI 2025*, pp. 1872-1878.
- [23] Srividya, K., et al. (2024) 'Societal implications of AI in Finance, Healthcare and in various industries', *ICPECTS 2024*, pp. 1-5.
- [24] Adamyk, O., et al. (2023) 'Trustworthy AI: A Fuzzy-Multiple Method for Evaluating Ethical Principles', *ACIT 2023*, pp. 608-613.
- [25] Deshpande, A. (2024) 'Regulatory compliance and AI: navigating challenges in finance', *ICKECS 2024*, pp. 1-5.
- [26] Pandow, B. A., et al. (2024) 'A review on AI-powered advancements in climate finance', *INDIACom 2024*, pp. 1409-1413.

- [27] Acharya, D. B., et al. (2025) 'Agentic AI: Autonomous Intelligence for Complex Goals', IEEE Access, 13, pp. 18912-18936.
- [28] Pangavhane, P., et al. (2023) 'Transforming finance through automation using AI-driven personal finance advisors', ICCAKM 2023, pp. 1-5.
- [29] Singh, G., et al. (2024) 'Understanding the Ethical Implications of Generative AI', ICESC 2024, pp. 1714-1719.
- [30] Al-Sinani, H.S., et al. (2025) 'Unleashing AI in ethical hacking', STM 2024, LNCS vol. 15235. Springer.