

Explainable Artificial Intelligence for Financial Fraud Detection: Bridging Transparency, Trust and Regulatory Compliance in Digital Banking Systems

M Ravi Kumar

Research Scholar
Department of Management
Osmania University, Hyderabad

Abstract:

The rapid growth of digital banking, mobile payments and online financial services has increased the speed, volume and complexity of financial transactions while simultaneously creating new opportunities for fraud. Highly accurate AI models, particularly ensemble and deep-learning models, can be difficult for customers, bank employees, auditors and regulators to understand. This creates a fundamental tension between predictive performance and explainability.

Explainable Artificial Intelligence (XAI) offers a potential solution by providing understandable explanations for why an AI system has classified a transaction as fraudulent or legitimate. Techniques such as SHAP (Shapley Additive Explanations), LIME, feature importance, counterfactual explanations and graphical explanations can help financial institutions establish greater transparency and accountability. This research examines how XAI can bridge the gap between fraud-detection accuracy, transparency, customer trust and regulatory compliance in digital banking. The analysis shows that AI adoption was already substantial in financial services, while explainability was becoming an important governance requirement. IBM's 2023 Global AI Adoption Index reported that approximately 42% of large organisations had actively deployed AI, while financial services was the industry most likely to report active AI deployment. In India, 59% of surveyed large enterprises had actively deployed AI, and 94% of Indian respondents considered the

Keywords: Explainable Artificial Intelligence, Financial Fraud Detection, Digital Banking, SHAP, LIME, Banking Security.

Introduction

Digital transformation has fundamentally changed the structure of the banking industry. Customers can now open accounts, transfer money, make payments, apply for financial products and manage investments through mobile applications and internet banking platforms. Although digital banking provides significant advantages in terms of accessibility, speed and convenience, it has also expanded the attack surface available to fraudsters. Fraud can occur through stolen credentials, account takeover,

phishing, identity theft, payment manipulation, card fraud, social engineering, money laundering and increasingly sophisticated online scams.

Traditional fraud-detection approaches generally rely on predefined rules. For example, a bank may flag a transaction when its value exceeds a predetermined threshold, when a customer logs in from an unusual geographical location or when several transactions occur within a short period. Rule-based systems are relatively easy to understand and audit. However, they have important limitations. Fraudsters can modify their behaviour to avoid predefined rules, while legitimate customers may also produce unusual transactions. Consequently, rule-based systems can generate large numbers of false positives, requiring employees to investigate transactions that are ultimately legitimate.

Machine learning provides a more flexible alternative. Instead of relying entirely on manually defined rules, machine-learning algorithms can learn patterns from historical transaction data. Classification algorithms, random forests, gradient boosting, neural networks, anomaly-detection methods and ensemble models can identify relationships among transaction amount, time, location, device characteristics, customer behaviour and transaction history.

The European Banking Authority's 2023 Risk Assessment Report illustrates the increasing use of machine-learning techniques in financial institutions. Among the techniques reported in the EBA material, decision trees/random forests were reported by 87%, regression analysis by 86%, natural-language processing by 81%, support-vector machines by 31%, and generative AI by 46% of respondents in the relevant survey.

This is the central role of Explainable Artificial Intelligence. XAI attempts to make AI outputs understandable to human users by identifying the factors that contributed to a prediction. For fraud detection, an XAI system could, for example, identify that a transaction was considered suspicious because it occurred from a new device, differed significantly from the customer's normal transaction amount and was associated with an unusual location. Such information allows a fraud analyst to investigate the case more efficiently.

Academic research has also begun to examine this issue. A 2023 study published in Finance Research Letters developed a user-centred XAI approach for financial fraud detection by combining an ensemble predictive model with Shapley-value-based explanations. The researchers specifically focused on connecting XAI explanations with the requirements of external stakeholders. Shapley-value approaches were among the most commonly reported methods, alongside documentation, graphical tools, feature-importance scores and sensitivity analysis.

Therefore, the challenge facing digital banks is no longer simply to ask, "Can AI detect fraud accurately?" The more important question is: "Can AI detect fraud accurately while providing sufficiently understandable, fair, auditable and trustworthy explanations?"

This research addresses that question by examining the relationship between XAI, financial fraud detection, transparency, customer and institutional trust, and regulatory compliance. It argues that the most effective future fraud-detection architecture is not one based exclusively on maximum predictive accuracy, but one that balances accuracy, explainability, privacy, fairness, human oversight and regulatory accountability.

Objectives

- To examine the role of Artificial Intelligence and Machine Learning in financial fraud detection.
- To evaluate the contribution of Explainable AI techniques, particularly SHAP, LIME, feature importance and counterfactual explanations, to understanding fraud-detection decisions.
- To analyse the relationship between explainability and trust among banking customers, fraud analysts, management and regulators.
- To develop recommendations for implementing an XAI-enabled fraud-detection framework that balances predictive accuracy, transparency, privacy, fairness and operational efficiency.

Statistical Analysis**Growth of digital fraud**

The FBI's 2023 IC3 statistics provide a useful indicator of the scale of online financial crime.

Indicator	2022	2023	Change
IC3 complaints	Approx. 800,000	880,418	+~10%
Reported potential losses	Approx. US\$10.3 bn	>US\$12.5 bn	+22%
Phishing/spoofing complaints	—	>298,000	Major category

The increase in potential losses was substantially greater than the increase in complaints. This suggests that the economic impact per reported complaint was increasing, reinforcing the need for sophisticated detection systems.

A simple ratio can illustrate:

2023 average potential loss per complaint

$\$12.5 \text{ billion} / 880,418 \approx \$14,198$

It provides an indication of the financial magnitude represented by each reported complaint.

AI adoption

IBM's 2023 Global AI Adoption Index found that:

- 42% of large organisations surveyed had actively deployed AI.
- 40% were actively exploring AI.
- Approximately half of IT professionals in financial services reported active AI deployment.
- 59% of surveyed large Indian organisations had actively deployed AI.

Therefore, AI was already moving from experimentation toward operational deployment.

The financial-services sector's relatively high adoption is particularly significant because fraud detection is a data-intensive application where AI can analyse large transaction volumes in near real time.

Importance of explainability

IBM's data provides strong evidence concerning the business importance of explainability.

Region	Respondents considering AI decision explainability important
Global enterprise sample	83%
India	94%
China	93%
UAE	93%
LATAM	89%

India's 94% figure is particularly relevant to digital banking research because India has experienced rapid growth in digital payments and AI adoption. It suggests that explainability is not simply a European or North American regulatory concern, it is also highly relevant to Indian enterprises.

The role of SHAP

SHAP is particularly useful because it can estimate the contribution of individual features to a prediction. For example:

Feature	SHAP contribution	Interpretation
New device	+0.28	Increases fraud probability
Unusual location	+0.21	Increases fraud probability
High transaction value	+0.18	Increases fraud probability
Unusual transaction time	+0.12	Increases fraud probability
Normal customer history	-0.08	Reduces fraud probability

The analyst can therefore see that the model did not simply classify the transaction as fraudulent arbitrarily. Instead, several measurable characteristics contributed to the decision.

This improves local interpretability, meaning that the explanation concerns an individual transaction.

SHAP can also contribute to global interpretability by showing which features are generally important across a large population of transactions.

LIME and counterfactual explanations

LIME can provide a simplified approximation of a complex model around an individual prediction. It can therefore answer:

"Which factors were most influential for this particular transaction?"

Counterfactual explanations answer a different but equally useful question:

"What would need to change for the prediction to become legitimate?"

Transparency and Trust

Trust is fundamental to banking. Customers expect financial institutions to protect their money while avoiding unnecessary restrictions on legitimate transactions.

A completely opaque fraud-detection system creates a communication problem. **Consider a customer**

whose transaction is declined:

Without XAI:

"Your transaction was blocked because our security system detected suspicious activity."

With appropriate XAI:

"This transaction was temporarily flagged because it differs substantially from your recent transaction behaviour and was initiated using a device not previously associated with your account. Please verify the transaction."

The second explanation provides considerably more useful information without exposing sensitive details about the bank's entire fraud-detection strategy.

This distinction is important because transparency does not mean revealing the complete algorithm. Excessive disclosure could allow criminals to reverse-engineer detection mechanisms.

Therefore, banks require controlled transparency.

XAI and Regulatory Compliance

The GDPR establishes important safeguards around automated individual decision-making. Individuals are provided with protection against certain decisions based solely on automated processing and identifies safeguards such as human intervention, the ability to express a viewpoint and the ability to contest a decision in applicable circumstances.

The GDPR is particularly relevant to banking because automated profiling can affect customers' access to financial services. Furthermore, GDPR provisions recognise fraud prevention as a legitimate context for certain automated processing, but require suitable safeguards. XAI can therefore support compliance by creating a documented logic trail around automated decisions.

Key Findings

- Fraud is a large-scale digital problem-The FBI reported more than 880,000 internet-crime complaints and over US\$12.5 billion in potential losses in 2023. The 22% year-on-year increase in reported losses demonstrates the economic significance of digital crime.
- AI adoption is already substantial-AI was no longer merely an experimental technology by 2023. IBM reported that 42% of large organisations had actively deployed AI and that financial services was among the leading sectors in adoption.
- Explainability is a business requirement-Globally, 83% of surveyed enterprise respondents considered the ability to explain AI decisions important. In India, the figure reached 94%.
- SHAP is particularly relevant to financial applications- EBA research indicates that Shapley-value techniques were among the most frequently used approaches for explaining machine-learning models in financial institutions.
- Explainability has limitations-An explanation does not automatically mean that a model is fair, accurate or compliant. A model can provide a technically convincing explanation while still being biased because of poor-quality training data.

Challenges of Implementing XAI

- Accuracy versus interpretability- Simple models such as logistic regression and decision trees can be easier to interpret but may not always capture complex fraud patterns as effectively as advanced ensemble or deep-learning models.
- Explanation fidelity-A simplified explanation may not perfectly represent the underlying model. Therefore, institutions should validate whether explanations faithfully represent model behaviour.
- Privacy-Fraud-detection models require large quantities of customer data. Financial institutions must ensure that the use of this data complies with privacy and security requirements.
- Adversarial behaviour-If fraudsters learn exactly which features trigger a fraud alert, they may deliberately modify their behaviour. Therefore, explanations provided to customers should be carefully designed.

Recommendations

Based on the findings, digital banks should:

- Adopt XAI as part of model governance rather than as an optional software feature.
- Use multiple explanation techniques, particularly SHAP, feature importance and counterfactual explanations, rather than depending on one method.
- Implement human oversight for high-risk decisions, particularly account blocking and decisions with significant customer consequences.
- Maintain comprehensive model documentation, including training data, model versions, performance metrics and explanation records.
- Monitor fairness and false-positive rates across customer groups.
- Use customer-facing explanations carefully, providing meaningful information without revealing sensitive fraud-detection rules.
- Regularly validate and recalibrate models because fraud patterns evolve rapidly.
- Integrate XAI with privacy and cybersecurity controls rather than treating transparency independently from data protection.
- Create audit trails that allow regulators and internal auditors to reconstruct how significant fraud decisions were reached.
- Combine AI with human expertise because fraud investigators can provide contextual knowledge that an automated model may not possess.

Conclusion

The rapid expansion of digital banking has created substantial opportunities for financial inclusion and convenience, but it has simultaneously increased exposure to sophisticated forms of financial and cyber fraud. The future of trustworthy AI-based financial fraud detection lies not in choosing between accuracy and explainability, but in designing systems that achieve an appropriate balance between predictive performance, transparency, customer trust and regulatory accountability.

For digital banking institutions, XAI should consequently become a core component of responsible AI governance rather than a secondary technical feature.

REFERENCES:

1. European Banking Authority. (2023). Follow-up report on machine learning for IRB models.
2. European Banking Authority. (2023). Risk Assessment Report: December 2023.
3. European Union. (2016). Regulation (EU) 2016/679: General Data Protection Regulation.
4. FBI Internet Crime Complaint Center. (2024). 2023 Internet Crime Report. Federal Bureau of Investigation.
5. IBM. (2024). IBM Global AI Adoption Index 2023. IBM Newsroom.
6. Molnar, C. (2022). Interpretable Machine Learning. Interpretable Machine Learning.
7. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 206–215.
8. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
9. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.
10. Černevičienė, J., & Kabašinskas, A. (2023). Applications of Explainable Artificial Intelligence in Finance—a systematic review of Finance, Information Systems, and Computer Science literature. *Management Review Quarterly*.