Pingale Neha Pravin ., 2026, 14:5
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Explainable Artificial Intelligence (XAI) In The Banking Sector: A Review Of Techniques For Transparent Credit, Fraud And Compliance Decisions

Pingale Neha Pravin[1]; Prof. M. R. Kumbhoje[2]; Prof. M. C. Kolse[3]; Prof. S. K. Said[4]
Department Of Artificial Intelligence And Data Science Jaihind College Of Engineering Pune, Maharashtra, India;

Abstract—Banks increasingly use complex machine learning models such as neural networks, random forests and gradient boosting for credit scoring, loan approval, fraud detection, risk assessment and anti-money laundering. These models process sensitive financial data and directly affect people's access to credit and other essential services, yet many operate as black boxes that cannot explain why a loan was rejected or a transaction flagged. This raises concerns about fairness, bias, accountability and compliance with regulations that require meaningful justification of automated decisions. This paper reviews Explainable Artificial Intelligence (XAI) for banking. It examines why explainability is needed, then surveys inherently interpretable models such as logistic regression and decision trees, post-hoc techniques including LIME, SHAP, accumulated local effects and feature importance, and counterfactual explanations that tell customers what would change a decision. Applications in credit risk, fraud detection and customer decision support are discussed, together with fairness metrics and regulatory frameworks including the GDPR, the EU AI Act, model risk management guidance and the Reserve Bank of India's responsible AI framework. A layered XAI architecture for banking is proposed, with research gaps, objectives and a comparison of techniques. The review concludes that explainability is essential for transparent, ethical and reliable AI-driven banking.

Index Terms—Explainable AI, banking, credit scoring, fraud detection, SHAP, LIME, counterfactual explanations, algorithmic fairness, regulatory compliance

I. INTRODUCTION

Machine learning now influences many banking decisions: whether an applicant receives a loan and at what interest rate, whether a card transaction is blocked as fraud, and which accounts are investigated for money laundering. Benchmarking by Lessmann *et al.* [1] showed that ensemble methods such as random forests and gradient boosting often outperform the traditional logistic regression scorecard in predicting credit default.

Greater accuracy, however, comes at the cost of transparency. A loan officer who cannot explain a rejection cannot satisfy the customer, an auditor cannot verify that the model is sound, and a regulator cannot confirm that it does not discriminate. The GDPR [2] restricts solely automated decisions with significant effects, and supervisory guidance on model risk management [3] requires banks to understand and validate their models. In India, the Reserve Bank's committee on responsible and ethical AI [4] has emphasized explainability and accountability for AI in financial services.

This paper reviews XAI for banking, proposes an architecture for explainable decision systems and identifies research gaps and challenges.

II. BACKGROUND AND RELATED WORK

1. Interpretability And Explainability

An interpretable model is one whose working is understandable by design, such as a short decision tree or a linear scorecard. Explainability refers to methods that describe the behaviour of any model, including complex ones. Rudin [5] argues that for high-stakes decisions, inherently interpretable models

© 2026 Pingale Neha Pravin This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

should be preferred whenever they achieve comparable accuracy.

2. Types Of Explanation

Explanations can be global, describing how a model behaves overall, or local, explaining one decision. They can be feature attributions, example-based or counterfactual. Fig. 1 classifies XAI techniques relevant to banking.

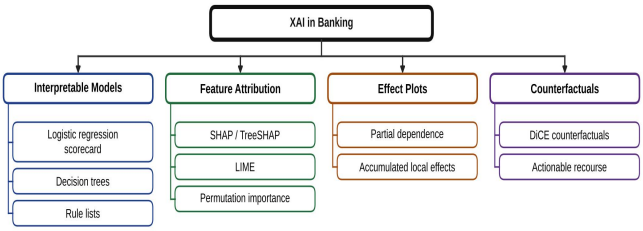


Fig. 1. Classification of XAI techniques for banking.

3. Review Of Related Studies

Ribeiro *et al.* [6] proposed LIME, which explains an individual prediction by perturbing the input and fitting a simple local model. Lundberg and Lee [7] introduced SHAP, which attributes the prediction to features using Shapley values from cooperative game theory and unifies several earlier methods; TreeSHAP [8] computes exact values efficiently for tree ensembles widely used in credit scoring.

Bussmann *et al.* [9] applied SHAP to an XGBoost credit-risk model for small and medium enterprises and showed how similarity among explanations can group borrowers into interpretable risk segments, supporting both accuracy and transparency. Apley and Zhu [10] proposed accumulated local effects plots, which describe how predictions change with a feature while avoiding the bias of partial dependence plots when features are correlated, as income and loan amount often are.

Wachter *et al.* [11] argued that counterfactual explanations, statements such as "the loan would have been approved if annual income were higher by a stated amount," can satisfy the spirit of GDPR without revealing model internals. Mothilal *et al.* [12] generated diverse and feasible counterfactuals, and Ustun *et al.* [13] introduced actionable recourse, ensuring that suggested changes are ones a customer can actually make.

On fairness, Hardt *et al.* [14] defined equality of opportunity, requiring equal true-positive rates across groups, providing a measurable fairness criterion for lending. Public datasets such as the Taiwan credit-card default data [15] and a European credit-card fraud dataset [16] are widely used to study these methods. Table I summarizes the surveyed literature.

Ref.	Author(s) / Year	Technique / Approach	Key Contribution	Limitation / Gap
[1]	Lessmann et al., 2015	Benchmark of credit-scoring classifiers	Ensembles outperform logistic regression	Accuracy focus; no explanations
[6]	Ribeiro et al., 2016	LIME local surrogate models	Model-agnostic local explanations	Instability across runs
[7]	Lundberg and Lee, 2017	SHAP Shapley-value attribution	Consistent, unified feature attribution	Costly for some models
[8]	Lundberg et al., 2020	TreeSHAP for tree ensembles	Exact, fast explanations for GBDT	Tree models only
[9]	Bussmann et al., 2021	SHAP on XGBoost credit model	Interpretable risk segments for SMEs	Single dataset
[11]	Wachter et al., 2018	Counterfactual explanations	Recourse without revealing model	May suggest infeasible changes
[12]	Mothilal et al., 2020	DiCE diverse counterfactuals	Multiple feasible alternatives	Computational cost

Ref.	Author(s) / Year	Technique / Approach	Key Contribution	Limitation / Gap
[13]	Ustun et al., 2019	Actionable recourse	Guarantees changeable features	Linear models
[5]	Rudin, 2019	Case for interpretable models	Interpretable models for high stakes	May lose some accuracy

TABLE I Summary of Literature Survey

III. RESEARCH GAP AND OBJECTIVES

1. Identified Research Gaps

The following gaps are identified:

- 1. Explanations from different methods can disagree, and there is no standard for choosing among them.
- 2. The stability and faithfulness of explanations are rarely validated in production banking systems.
- 3. Explanations designed for data scientists are often unsuitable for customers or branch staff.
- 4. Fraud and anti-money laundering models raise explanation challenges distinct from credit scoring, which are less studied.
- 5. Practical guidance aligning XAI outputs with Indian regulatory expectations is still emerging.

2. Objectives Of The Review

The objectives of this review are:

• To explain the need for explainability in banking AI.

• To review interpretable models and post-hoc XAI techniques, including counterfactuals.

• To examine applications in credit risk, fraud detection and customer decision support.

• To propose a layered XAI architecture for banking decision systems.

• To analyse challenges such as the accuracy–interpretability trade-off, privacy and regulatory compliance.

IV. DESIGN OF THE PROPOSED SYSTEM

The architecture in Fig. 2 makes explanation and governance integral parts of the banking decision pipeline.

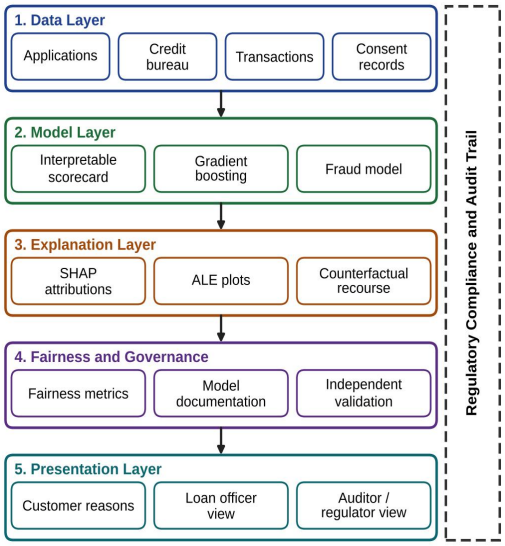


Fig. 2. Proposed layered XAI architecture for banking decision systems.

1. Data Layer

Application data, credit bureau records, transaction histories and account information are collected with consent, validated and secured.

2. Model Layer

Both an interpretable model, such as a scorecard, and a high-performance model, such as gradient boosting, are trained. The simpler model serves as a benchmark and fallback.

3. Explanation Layer

SHAP values provide local and global feature attributions, accumulated local effects describe feature effects, and counterfactual generators produce actionable recourse statements for declined customers.

4. Fairness And Governance Layer

Fairness metrics are computed across protected groups, model documentation records data,

assumptions and limitations, and independent validators review models before deployment.

5. Presentation Layer

Different views serve different users: customers receive plain-language reasons and steps for improvement, loan officers see the main factors with supporting data, and auditors and regulators access full documentation and logs.

6. Methodology

The workflow is shown in Fig. 3. The SHAP value of feature i for prediction f(x) is the average marginal contribution of the feature over all subsets S of the other features F:

$$\varphi_i = \Sigma_S\ w(S) \cdot [\ f(S\ \cup \{i\}) - f(S)\ ] \quad (1)$$

A counterfactual x′ for applicant x is obtained by solving

$$x' = argmin\ d(x, x')\ \ s.t.\ \ f(x') = approve \quad (2)$$

using only features the customer can change. Fairness under equality of opportunity is checked by comparing true-positive rates across groups.

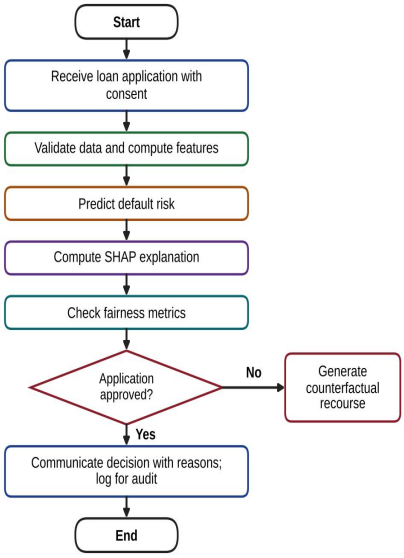


Fig. 3. Explainable loan decision workflow.

V. DISCUSSION AND COMPARISON

Table II compares explanation techniques. Interpretable models offer direct transparency; LIME is flexible but can be unstable; SHAP is consistent and widely adopted; ALE handles correlated features; and counterfactuals are the most meaningful for customers.

Technique	Scope	Stability	Best For
Scorecard / tree	Global	High	Regulated credit decisions
LIME	Local	Low–Medium	Quick local insight
SHAP / TreeSHAP	Local + global	High	Credit and fraud models
ALE plots	Global	High	Correlated features
Counterfactuals	Local	Medium	Customer recourse

TABLE II Comparison of XAI Techniques for Banking

VI. CHALLENGES AND FUTURE DIRECTIONS

Table III summarizes the main challenges. Future directions include standardized explanation quality metrics, explanations for graph-based fraud and anti-money-laundering models, privacy-preserving explanations that do not reveal sensitive training data, natural-language explanations generated by language models under strict controls, and compliance with risk-based regulations such as the EU AI Act [17], which classifies credit scoring as high-risk.

Challenge	Mitigation
Accuracy vs. interpretability	Benchmark against interpretable model
Inconsistent explanations	Use stable methods; validate faithfulness
Bias and discrimination	Fairness metrics; bias mitigation
Data privacy	Aggregate explanations; access controls
Regulatory compliance	Documentation aligned to RBI / GDPR / AI Act

TABLE III Challenges and Mitigation

VII. CONCLUSIONS

As banks rely more on AI for decisions that affect people's financial lives, explainability has become a

necessity rather than an option. The reviewed literature provides a mature toolkit: interpretable models where possible, SHAP and related methods for complex models, and counterfactual explanations that give customers actionable recourse. Combined with fairness testing, documentation and audience-specific presentation, XAI builds customer trust, helps detect and reduce bias, supports regulatory compliance and enables safer adoption of AI in financial services.

1. Acknowledgment

The authors express their sincere gratitude to the Department of Artificial Intelligence and Data Science, Jaihind College of Engineering, Pune, for providing the facilities, guidance and encouragement needed to carry out this work. The authors also thank the faculty members and colleagues whose valuable suggestions helped improve the quality of this review.

REFERENCES

1. S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," Eur. J. Oper. Res., vol. 247, no. 1, pp. 124–136, 2015.
2. European Parliament and Council, "Regulation (EU) 2016/679 (General Data Protection Regulation)," Off. J. Eur. Union, L 119, 2016.
3. Board of Governors of the Federal Reserve System, "Supervisory guidance on model risk management," SR Letter 11-7, 2011.
4. Reserve Bank of India, "Report of the Committee to develop a Framework for Responsible and Ethical Enablement of Artificial Intelligence (FREE-AI)," RBI, Mumbai, Aug. 2025.
5. C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nat. Mach. Intell., vol. 1, no. 5, pp. 206–215, 2019.
6. M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?": Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD, 2016, pp. 1135–1144.
7. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. NeurIPS, vol. 30, 2017.
8. S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," Nat. Mach. Intell., vol. 2, no. 1, pp. 56–67, 2020.
9. N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable machine learning in credit risk management," Comput. Econ., vol. 57, no. 1, pp. 203–216, 2021.
10. D. W. Apley and J. Zhu, "Visualizing the effects of predictor variables in black box supervised learning models," J. R. Stat. Soc. Ser. B, vol. 82, no. 4, pp. 1059–1086, 2020.
11. S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," Harvard J. Law Technol., vol. 31, no. 2, pp. 841–887, 2018.
12. R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in Proc. ACM Conf. Fairness, Accountability, Transparency (FAT)*, 2020, pp. 607–617.
13. B. Ustun, A. Spangher, and Y. Liu, "Actionable recourse in linear classification," in Proc. ACM FAT, 2019, pp. 10–19
14. M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in Proc. NeurIPS, 2016, pp. 3315–3323.
15. I.-C. Yeh and C.-H. Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," Expert Syst. Appl., vol. 36, no. 2, pp. 2473–2480, 2009.
16. A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, "Calibrating probability with undersampling for unbalanced classification," in Proc. IEEE Symp. Ser. Comput. Intell. (SSCI), 2015, pp. 159–166.
17. European Parliament and Council, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," Off. J. Eur. Union, 2024.