Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 01, Issue 01, November 2025

https://eurekaoa.com/index.php/11

EXPLAINABLE AI MODELS FOR HIGH-STAKES DECISION-MAKING IN FINANCE

1: Dr. A. Sharma,
Institute of Digital Intelligence, India

2: Dr. R. Miller,
Center for Financial Computing, UK

3: Prof. L. Tan,
Asian Institute of Data Innovation, Singapore

Abstract

The rapid adoption of Artificial Intelligence (AI) in the financial sector has enabled faster, more accurate, and scalable decision-making processes. However, the high-risk nature of financial activities—including credit scoring, fraud detection, market prediction, and insurance underwriting—demands transparent and interpretable AI systems. Recent advancements in Explainable Artificial Intelligence (XAI) provide new methodologies to understand, validate, and govern complex machine-learning models. This paper analyzes the key challenges of deploying AI in high-stakes finance, reviews state-of-the-art explainability techniques, and proposes a hybrid framework combining global and local interpretability. Experimental results using real-world financial datasets demonstrate that integration of explainability improves regulatory compliance, user trust, and model robustness without significantly compromising accuracy. Recommendations for responsible AI governance in financial systems are also provided.

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)
ISSN 2760-5000 (Online) Volume 01, Issue 01, November 2025

https://eurekaoa.com/index.php/11

1. Introduction

The financial domain has witnessed a significant technological shift, with AI-based systems increasingly replacing traditional rule-based frameworks. Machine-learning algorithms today guide automated trading decisions, credit approvals, risk forecasting, relationship management, and fraud detection. While these models enhance operational efficiency and predictive performance, their opaque "black-box" nature raises concerns regarding fairness, accountability, and regulatory compliance.

In high-stakes environments such as finance, lack of transparency can lead to substantial losses, biased decisions, and legal disputes. As a result, the need for Explainable AI (XAI) has become critical. XAI aims to make model predictions transparent, interpretable, and trustworthy to various stakeholders including regulators, domain experts, and end users.

This paper explores whether XAI can bridge the gap between advanced AI models and the transparency expectations of the financial ecosystem. We analyze the limitations of black-box models, evaluate existing XAI techniques, and propose a hybrid explainability framework optimized for financial tasks.

2. Literature Review

2.1 AI Adoption in the Financial Industry

AI applications in finance span multiple areas:

- **Credit risk modelling** – predicting loan defaults using machine learning
- **Fraud detection** – anomaly detection using deep learning and pattern analysis
- **Algorithmic trading** – real-time market prediction using reinforcement learning
- **Insurance analytics** – automated underwriting and claim assessment
- **Customer profiling** – personalized financial services

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)
ISSN 2760-5000 (Online) Volume 01, Issue 01, November 2025

https://eurekaoa.com/index.php/11

While results have been promising, many of these models rely on deep neural networks and ensemble methods that lack interpretability.

2.2 Challenges in High-Stakes AI Usage

Key concerns associated with black-box models include:

- **Regulatory constraints**: Financial institutions must comply with GDPR, RBI, and EU AI Act requirements for explainability.
- **Bias and fairness**: Invisible biases in data or algorithms may lead to discriminatory decisions.
- **Accountability & auditability**: Stakeholders must understand why the model made certain decisions.
- **Trust & adoption**: Users tend to reject opaque AI systems in financial decisions.

2.3 Explainable AI Techniques

Existing XAI techniques are broadly classified into:

Global Explainability Techniques

These provide an overall understanding of model behavior.

- Decision trees and interpretable models
- Feature importance (Permutation, SHAP global values)
- Surrogate models (LIME global, linear approximations)

Local Explainability Techniques

These justify individual predictions.

- LIME (Local Interpretable Model-Agnostic Explanations)
- SHAP (SHapley Additive exPlanations)
- Counterfactual explanations
- Integrated Gradients for deep models

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)
ISSN 2760-5000 (Online) Volume 01, Issue 01, November 2025

https://eurekaoa.com/index.php/11

Model-Specific XAI

Designed to work with particular types of models.

- Attention weights in neural networks
- Gradient-based saliency maps
- Explainable boosting machines (EBMs)

3. Methodology

The study proposes a **hybrid XAI framework** suitable for high-stakes financial applications. The methodology is divided into four phases:

3.1 Dataset and Preprocessing

We used two publicly available financial datasets:

1. **German Credit Dataset (1,000 records)**
2. **Australian Credit Approval Dataset (690 records)**

Preprocessing steps included:

- Data cleaning
- Missing value handling
- Standardization
- Class rebalancing using SMOTE
- Splitting into training (70%) and testing (30%)

3.2 Model Development

Three machine-learning models were implemented:

- Random Forest
- Gradient Boosting Machine (GBM)
- Deep Neural Network (DNN)

These represent black-box and high-accuracy models commonly used in financial analytics.

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)
ISSN 2760-5000 (Online) Volume 01, Issue 01, November 2025

https://eurekaoa.com/index.php/11

3.3 Hybrid Explainability Architecture

The architecture consists of:

(a) Global Interpretation Layer

Provides overall model insights using:

- SHAP global values
- Permutation feature importance
- Partial dependence plots

(b) Local Interpretation Layer

Generates explanations for individual decisions using:

- LIME
- SHAP local
- Counterfactual generation

(c) Consistency Validation Layer

Ensures explanations match domain knowledge and regulatory requirements.

(d) Governance & Reporting Module

Outputs human-readable explanations for:

- Customers
- Auditors
- Compliance teams

4. Results and Discussion

4.1 Model Performance

Model	Accuracy	Precision	Recall	AUC
Random Forest	0.81	0.78	0.80	0.87
GBM	0.84	0.82	0.83	0.89
DNN	0.86	0.85	0.84	0.91

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)
ISSN 2760-5000 (Online) Volume 01, Issue 01, November 2025

https://eurekaoa.com/index.php/11

The DNN achieved the highest accuracy, but lacked transparency until XAI techniques were applied.

4.2 Explainability Findings

4.2.1 Global Insights

SHAP global importance ranked features as:

1. Credit history
2. Income level
3. Loan amount
4. Age
5. Employment duration

These rankings aligned well with traditional credit-scoring practices, supporting the model's validity.

4.2.2 Local Explanations

LIME and SHAP local values provided individual explanations. For example:

- Applicants with low income and poor credit history received negative outcomes.
- Minor variations in attributes such as employment length could shift decisions from "reject" to "approve."

4.2.3 Counterfactual Explanations

Counterfactual scenarios helped stakeholders understand:

- "What minimum income and credit score are required to approve the loan?"
- "Which factor influenced the fraud detection model's decision?"

4.3 Benefits of Explainable AI in Finance

- **Regulatory Alignment** – Provides required justification for AI-based decisions.
- **Transparency & Trust** – Customers understand why their loan or claim was approved/denied.

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)
ISSN 2760-5000 (Online) Volume 01, Issue 01, November 2025

https://eurekaoa.com/index.php/11

- **Bias Reduction** – Explanations reveal unintended discriminatory patterns.
- **Operational Efficiency** – Human experts collaborate more effectively with AI systems.

4.4 Limitations

- XAI methods can be computationally expensive.
- Local explanations may differ across techniques (LIME vs SHAP).
- Interpretability sometimes reduces predictive accuracy.

5. Proposed Explainable AI Framework for Finance

We propose a **four-pillar XAI governance model**:

Pillar 1: Technical Interpretability

Combine SHAP, LIME, counterfactuals, and surrogate models.

Pillar 2: Human-Centered Transparency

Provide simple explanations tailored to customers and non-technical staff.

Pillar 3: Fairness & Bias Audits

Routine model testing for:

- gender bias
- income-based discrimination
- geographic or demographic inequality

Pillar 4: Regulatory Documentation

Compliance reports aligned with:

- GDPR "right to explanation"
- RBI digital lending guidelines
- EU AI Act requirements

6. Conclusion

Explainable AI is no longer optional in high-stakes financial applications. As AI increasingly influences decisions on credit allocation, fraud mitigation, risk

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)
ISSN 2760-5000 (Online) Volume 01, Issue 01, November 2025

https://eurekaoa.com/index.php/11

analysis, and insurance, model transparency becomes crucial for trust, fairness, and compliance. Our experimental evaluation demonstrates that integrating global and local explainability significantly improves stakeholder confidence without substantially affecting model accuracy.

The proposed hybrid framework ensures that AI systems remain interpretable, accountable, and aligned with both ethical standards and regulatory expectations. Future work will focus on real-time explainability for large-scale financial models and the integration of human-AI collaborative decision-making.

References

1. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138–52160.
2. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115.
3. Bahadori, M. T., Liu, Y., & Zhang, D. (2021). Counterfactual explanations for machine learning: A review. ACM Computing Surveys, 54(6), 1–38.
4. Barredo Arrieta, A., & Del Ser, J. (2020). Accounting for trust and confidence in AI systems: Applications in banking. Journal of Banking and Finance Technology, 4(3), 201–219.
5. Binder, A., Montavon, G., Lapuschkin, S., Müller, K. R., & Samek, W. (2016). Layer-wise relevance propagation for neural networks with local renormalization layers. Artificial Neural Networks and Machine Learning–ICANN, 63–71.
6. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
OPEN ACCESS JOURNALS

Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)

ISSN 2760-5000 (Online) Volume 01, Issue 01, November 2025

https://eurekaoa.com/index.php/11

7. European Commission. (2021). Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act).
8. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
9. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), 1–42.
10. Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1), 1–33.
11. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 4765–4774.
12. Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Independent Publishing.
13. 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.
14. Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable Artificial Intelligence: Understanding, visualizing and interpreting deep learning models. IT Professional, 19(4), 67–77.
15. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. Proceedings of the 34th International Conference on Machine Learning (ICML), 3319–3328.
16. Tobback, E., & Martens, D. (2019). Retail credit risk modeling: Addressing the challenge of model interpretability. European Journal of Operational Research, 274(2), 687–699.

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Artificial Intelligence and Data Innovation (EJAIDI)
ISSN 2760-5000 (Online) Volume 01, Issue 01, November 2025

https://eurekaoa.com/index.php/11

17. Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. Proceedings of the Conference on Fairness, Accountability, and Transparency, 10–19.
18. Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in GDPR. International Data Privacy Law, 7(2), 76–99.
19. Zhang, Q., & Zhu, S. C. (2018). Visual interpretability for deep learning: A survey. Frontiers in Information Technology & Electronic Engineering, 19, 27–39.