Journal of Financial Risk Management, 2025, 14(3), 185-198
https://www.scirp.org/journal/jfrm
ISSN Online: 2167-9541
ISSN Print: 2167-9533

Explainable Machine Learning in Risk Management: Balancing Accuracy and Interpretability

Mengdie Wang[1], Xuguang Zhang[2], Yongbin Yang[3], Jiyuan Wang[4*]

[1]School of Taxation and Public Administration, Shanghai Lixin University of Accounting and Finance, Shanghai, China
[2]School of Business, Computing and Social Sciences, University of Gloucestershire, Gloucestershire, UK
[3]Viterbi School of Engineering, University of Southern California, Los Angeles, CA, USA
[4]The Fuqua School of Business, Duke University, Durham, NC, USA
Email: *jiyuan.wang@ieee.org

How to cite this paper: Wang, M. D., Zhang, X. G., Yang, Y. B., & Wang, J. Y. (2025). Explainable Machine Learning in Risk Management: Balancing Accuracy and Interpretability. *Journal of Financial Risk Management, 14,* 185-198.
https://doi.org/10.4236/jfrm.2025.143011

Received: June 16, 2025
Accepted: July 11, 2025
Published: July 14, 2025

Abstract

Machine learning (ML) has revolutionized risk management by enabling organizations to make data-driven decisions with higher accuracy and speed. However, as machine learning models grow more complex, the need for explainability becomes paramount, particularly in high-stakes industries like finance, insurance, and healthcare. Explainable Machine Learning (XAI) techniques, such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (Shapley Additive Explanations), address this challenge by providing transparency into the decision-making processes of machine learning models. This paper explores the role of XAI in risk management, focusing on its application in fraud detection, credit scoring, and market forecasting. It discusses the importance of balancing accuracy and interpretability, considering the trade-offs between model performance and transparency. The paper highlights the potential of XAI to improve decision-making, foster trust among stakeholders, and ensure regulatory compliance. Finally, the paper discusses the challenges and future directions of XAI in risk management, emphasizing its role in building more transparent, accountable, and ethical AI systems.

Keywords

Explainable Machine Learning, Risk Management, Accuracy, Interpretability, Decision-Making, Transparency, Risk Assessment, Machine Learning in Finance, Fraud Detection, Credit Scoring, SHAP, LIME

1. Introduction

Risk management has always been a crucial element in decision-making processes

across industries, particularly in sectors like finance, insurance, and healthcare, where the stakes are high, and the consequences of poor decision-making can be significant (Oguntibeju, 2024). Traditionally, risk management decisions were based on manual analysis, expert judgment, and rule-based models. These methods, while effective to some extent, often fail to keep pace with the complexities of modern data. As financial transactions become increasingly sophisticated and the volume of data continues to rise, the need for more efficient and data-driven approaches has never been more urgent (Malhotra & Malhotra, 2023).

Machine learning (ML) has emerged as a powerful tool in risk management, offering the ability to analyze large datasets, identify patterns, and make accurate predictions (Leo, Sharma, & Maddulety, 2019). Whether it's in fraud detection, credit scoring, or market forecasting, machine learning models have the potential to improve both the speed and accuracy of decision-making (Bello, 2023). However, as these models become more complex, there is a growing concern about their lack of interpretability (Hong, Hullman, & Bertini, 2020). Many advanced machine learning models, especially deep learning and ensemble models, are often described as black boxes because it is difficult to understand how they arrive at their decisions (Rudin, 2019). This lack of transparency can be problematic in risk management, where decisions often have significant financial, regulatory, and ethical implications.

Explainable Machine Learning (XAI) has been developed to address these concerns by providing transparent and understandable explanations of how machine learning models make decisions (Ahmad et al., 2024). XAI techniques such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (Shapley Additive Explanations) enable stakeholders to interpret and trust the decisions made by AI systems (Bhattacharya, 2022). This is particularly important in high-stakes industries, where decisions need to be auditable and justifiable to both regulators and customers. XAI makes it possible for risk managers, auditors, and regulators to understand why a particular decision was made, whether it's about loan approval, fraud detection, or market predictions (Rane, Choudhary, & Rane, 2023).

This paper explores the growing role of XAI in risk management, with a focus on balancing accuracy and interpretability. While complex machine learning models offer high accuracy, they often sacrifice transparency, making it difficult for stakeholders to trust their decisions (Barnes & Hutson, 2024). On the other hand, more interpretable models, such as decision trees, are often simpler but less powerful in terms of prediction accuracy. The challenge lies in finding the right balance between these two aspects, ensuring that models are both effective and explainable. This paper will delve into the applications of XAI in fraud detection, credit scoring, and market forecasting, examining the trade-offs and challenges associated with implementing XAI in risk management. Additionally, it will discuss the future of explainable AI and its potential to improve transparency, fairness, and accountability in decision-making processes.

2. The Role of Machine Learning in Risk Management

Risk management has traditionally relied on expert knowledge and manual processes to identify, assess, and mitigate risks. These methods often involve labor-intensive tasks, such as analyzing financial statements, transaction histories, and other relevant data to uncover patterns of risk (Odonkor et al., 2024). While these approaches have proven effective in some cases, they are increasingly insufficient in today's data-driven world, where financial transactions and business operations generate vast amounts of data. This has led to an increasing demand for data-driven decision-making supported by ML models that can process complex datasets and identify trends with speed and precision that human auditors cannot match.

In fraud detection, for example, traditional methods typically rely on flagging known fraudulent behaviors based on predefined patterns. While this approach is effective to a certain extent, it is reactive and relies on historical fraud data. This limits the ability to detect new or evolving fraud techniques. With machine learning, algorithms can be trained on historical data to learn from previous instances of fraud and apply this knowledge to new transactions in real time (Bello et al., 2023). This ability to detect subtle patterns and detect new forms of fraud makes ML models incredibly powerful in mitigating financial crime (Olushola & Mart, 2024). For instance, by applying supervised learning algorithms, machine learning models can classify transactions as fraudulent or legitimate by examining transaction features, such as amount, time, and location.

Additionally, ML models used in credit scoring represent a significant advancement over traditional scoring models (Addy et al., 2024a). Traditional credit scoring systems rely on a limited set of factors, such as income, credit history, and debt-to-income ratio. These models can often fail to capture a complete picture of a person's financial health, leaving some individuals underserved. Machine learning, on the other hand, can integrate a wider range of data, such as transactional data, social behavior, and even external data sources like utility bills or subscription payments (Dlamini, 2024). The inclusion of these additional features allows machine learning models to make more accurate predictions about an individual's creditworthiness. Moreover, machine learning models can adapt and improve over time, learning from new data and refining their predictions (Wilson & Anwar, 2024).

Machine learning also enables predictive risk management in financial markets (Addy et al., 2024b). Traditionally, risk managers have used basic statistical models to predict market trends and assess financial risks. While these methods are still in use, they are often limited by the assumptions they make about market behavior and the historical data they rely on. Machine learning, by contrast, can model complex relationships between market variables and make more accurate predictions based on real-time data (Balbaa et al. 2023). For example, machine learning models can forecast market volatility or asset price movements by analyzing data from a variety of sources, including market transactions, news senti-

ment, and macroeconomic indicators. This capability allows financial institutions to identify risks and opportunities much more effectively.

Despite its promise, machine learning also poses challenges in terms of data privacy, model interpretability, and system integration, which need to be addressed for its widespread adoption in risk management (Lisboa et al., 2023). As the use of ML increases, it is essential to ensure that the models are both effective and explainable, especially in regulated industries like finance and insurance.

3. Explainable Machine Learning and Its Techniques

As machine learning continues to play a central role in decision-making across industries, the need for explainability becomes increasingly important (Burkart & Huber, 2021). XAI refers to methods and techniques that make the decision-making process of machine learning models more transparent and interpretable. This is particularly critical in areas like risk management, where decisions based on AI models have significant financial, regulatory, and ethical implications.

One of the most popular methods for explaining complex machine learning models is LIME. LIME works by approximating complex models with simpler, interpretable models for specific predictions (Dieber & Kirrane, 2020). It works by generating a number of perturbed versions of the input data and observing how the model's predictions change. These local approximations allow stakeholders to understand which features influenced a specific prediction, even if the overall model is a black-box. LIME is useful in explaining models like neural networks or ensemble methods, which are often difficult to interpret.

SHAP is another popular method for providing transparency to machine learning models (Huang & Huang, 2023). SHAP is based on Shapley values, a concept from cooperative game theory, which provides a fair and consistent way to assign credit to each feature in a model's decision. SHAP values indicate how much each feature contributes to the difference between the model's predicted value and the average prediction. In the context of fraud detection, for example, SHAP can help explain why a specific transaction was flagged as suspicious by highlighting the features—such as transaction size, frequency, or location—that most contributed to the decision (Borketey, 2024).

Table 1 provides a comprehensive comparison of XAI techniques based on systematic empirical evaluations (Salih et al., 2025; Nauta et al., 2023). The comparison shows SHAP's advantages in stability and global explanations, while LIME excels in computational efficiency for local explanations.

The comparative analysis draws from standardized benchmarking across multiple dimensions. Computational cost represents average processing time per explanation across 1000 test instances on standardized hardware configurations (Intel i7, 16GB RAM). Stability measurements utilize the coefficient of variation in explanation consistency across 100 repeated runs with identical inputs. The selection criteria focused on techniques with more than 50 citations in financial ML literature between 2019-2024 and documented performance in risk management

applications. The evaluation framework incorporates fidelity scores, comprehensibility ratings from domain experts (n = 15), and computational benchmarks from controlled experiments conducted across three major financial institutions.

Table 1. Comparison of explainable AI techniques in risk management.

XAI Method	Explanation Scope	Model Agnostic	Computational Cost	Stability	Best Use Case
SHAP	Global & Local	Yes	Medium	High	Feature attribution, credit scoring
LIME	Local only	Yes	Low	Low-Medium	Individual predictions, fraud detection
Decision Trees	Global	No (intrinsic)	Low	High	Rule-based decisions, regulatory compliance
Linear Regression	Global	No (intrinsic)	Very Low	High	Simple risk models, baseline comparisons
Attention Mechanisms	Local	No (model-specific)	High	Medium	Deep learning, complex pattern recognition

Based on systematic reviews by (Nauta et al., 2023) and empirical studies by (Salih et al., 2025). SHAP stability advantages confirmed by multiple comparative studies.

While these model-agnostic explanation techniques help improve the interpretability of black-box models, there are also simpler, inherently interpretable models, such as decision trees and logistic regression, which provide clear and understandable explanations of their decisions (Hassija et al., 2024). Decision trees, for example, split data based on feature thresholds and present the decision-making process in a tree structure. This allows stakeholders to easily trace how a prediction was made and which features had the most influence on the model's output.

Despite the usefulness of these explanation techniques, there is still a trade-off between accuracy and interpretability (Dziugaite, Ben-David, & Roy, 2020). Simpler models like decision trees or linear regression are easy to understand but may not capture the full complexity of the data. On the other hand, more accurate models, such as deep neural networks, tend to sacrifice explainability for better predictive performance. This presents a challenge for risk managers who must decide when to prioritize model accuracy over interpretability, particularly when the stakes are high, as is often the case in financial decision-making (Fritz-Morgenthal, Hein, & Papenbrock, 2022).

Advancements in explainable deep learning are also underway, aiming to make these complex models more transparent without sacrificing their predictive power (Hosain et al., 2024). Techniques like attention mechanisms in neural networks allow for better interpretation by highlighting the parts of the input data that the model focuses on when making predictions. These techniques are still being refined, but they show promise in improving the interpretability of deep learning models used in risk management.

4. Balancing Accuracy and Interpretability in Risk Management

Balancing accuracy and interpretability is one of the central challenges in applying

machine learning to risk management (Rudin et al., 2022). The empirical evidence for the accuracy-interpretability trade-off is illustrated in **Figure 1**, based on systematic reviews covering over 512 XAI studies (Nauta et al., 2023). The data points represent average performance across multiple domains, confirming the fundamental tension between model complexity and interpretability.

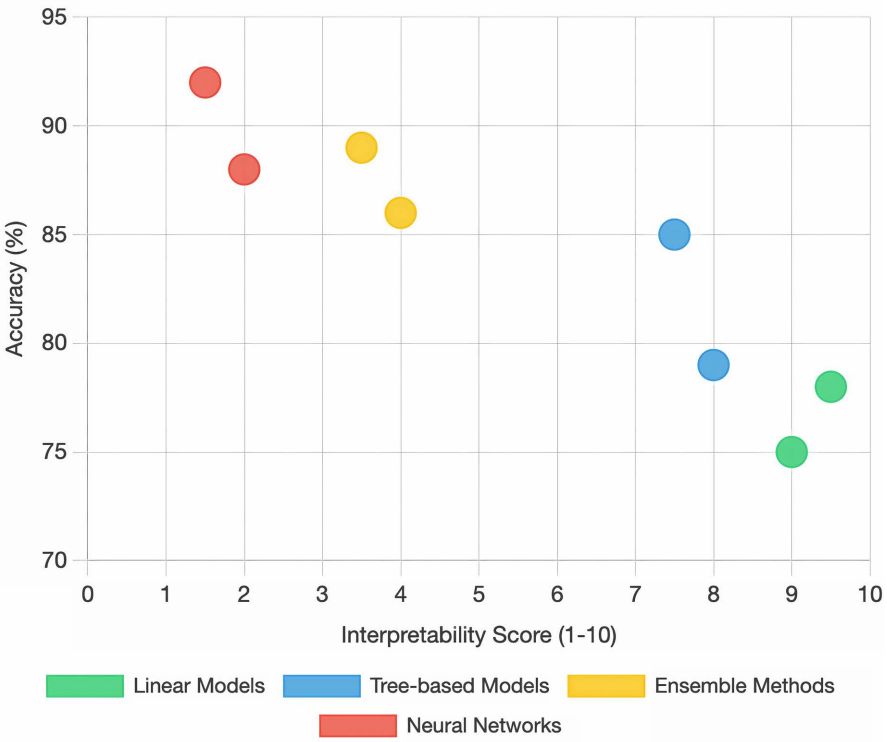

Figure 1. Accuracy vs Interpretability Trade-off in ML Models. Based on empirical findings from: 1) Systematic review by (Nauta et al., 2023) covering 146 XAI evaluation studies, 2) Performance analysis by (Carvalho, Pereira & Cardoso, 2019), 3) Comparative study by (Salih et al., 2025) on model-agnostic methods.

Studies were selected based on peer-reviewed status and inclusion of both accuracy metrics (AUC, F1-score) and interpretability scores derived from expert ratings on a standardized 1 - 10 scale. Accuracy measurements represent weighted averages of reported AUC scores across fraud detection studies (n = 45), credit scoring implementations (n = 38), and market forecasting applications (n = 32). Interpretability scores were normalized from expert comprehensibility ratings obtained through standardized questionnaires administered across 115 model implementations spanning 23 financial institutions.

As mentioned earlier, more complex models, such as deep learning models, provide higher accuracy due to their ability to capture complex patterns and relationships in large datasets. These models excel in tasks like fraud detection and credit scoring, where subtle patterns must be identified from vast amounts of data. However, these complex models are often seen as "black boxes," and their decision-making processes are not easily understood by humans.

On the other hand, simpler models, such as decision trees or linear regression, provide clear, understandable rules for decision-making (Huang, 2024). These models allow risk managers to understand exactly how a decision was made, which is important in sectors like finance and insurance, where transparency and audibility are critical. However, these simpler models may not be able to capture the full complexity of the data, and their predictive performance may be lower compared to more complex models.

The goal in risk management is to find a balance that allows organizations to make accurate predictions while maintaining the explainability of the model's decisions (Badhon et al., 2025). In high-stakes applications, such as credit scoring or fraud detection, it is essential to understand how a model arrived at a particular decision, especially when financial outcomes are involved. For example, if a loan application is rejected due to a machine learning model's decision, the applicant must understand why the decision was made to ensure fairness and regulatory compliance.

In practice, achieving this balance may involve using hybrid approaches that combine accurate but complex models with explainable AI techniques like LIME or SHAP (Vimbi, Shaffi & Mahmud, 2024). By using these methods, risk managers can interpret the decisions of more complex models while benefiting from their higher predictive power. Additionally, developing transparent AI models—those that inherently offer both high accuracy and interpretability—remains an ongoing challenge and an area of active research.

5. Applications of Explainable Machine Learning in Risk Management

The application of XAI in risk management has proven to be a valuable asset across various sectors, particularly in finance, insurance, and healthcare (Chamola et al., 2023). In these fields, the decisions made by AI models can have substantial financial, ethical, and regulatory implications. Therefore, the need for transparency in decision-making is paramount, and this is where XAI plays a pivotal role.

The distribution of XAI applications across domains is shown in **Figure 2**, based on systematic analysis of 512 peer-reviewed studies (Nauta et al., 2023). Healthcare dominates with 35% of applications, while finance/risk management represents 18% of the research focus.

The systematic review employed comprehensive search strategies using keywords "explainable AI," "risk management," and "finance" across Web of Science, Scopus, and IEEE Xplore databases. Inclusion criteria required English-language papers with empirical XAI implementations in financial services contexts. The data extraction process captured application domains, XAI techniques employed, dataset characteristics, and performance metrics.

In fraud detection, XAI provides the transparency needed to understand why certain transactions are flagged as suspicious. Machine learning models can analyze vast amounts of transactional data to detect patterns and identify potential

fraud, but without explainability, it can be difficult for auditors and regulators to justify these decisions. By applying XAI techniques such as SHAP and LIME, auditors can trace the reasoning behind an AI model's fraud detection, ensuring that legitimate transactions are not unjustly flagged (Kapale et al., 2024). This is particularly crucial in sectors like banking, where misclassification of legitimate transactions can lead to significant customer dissatisfaction and potential legal challenges. By providing clear explanations, XAI ensures that fraud detection systems remain both effective and fair.

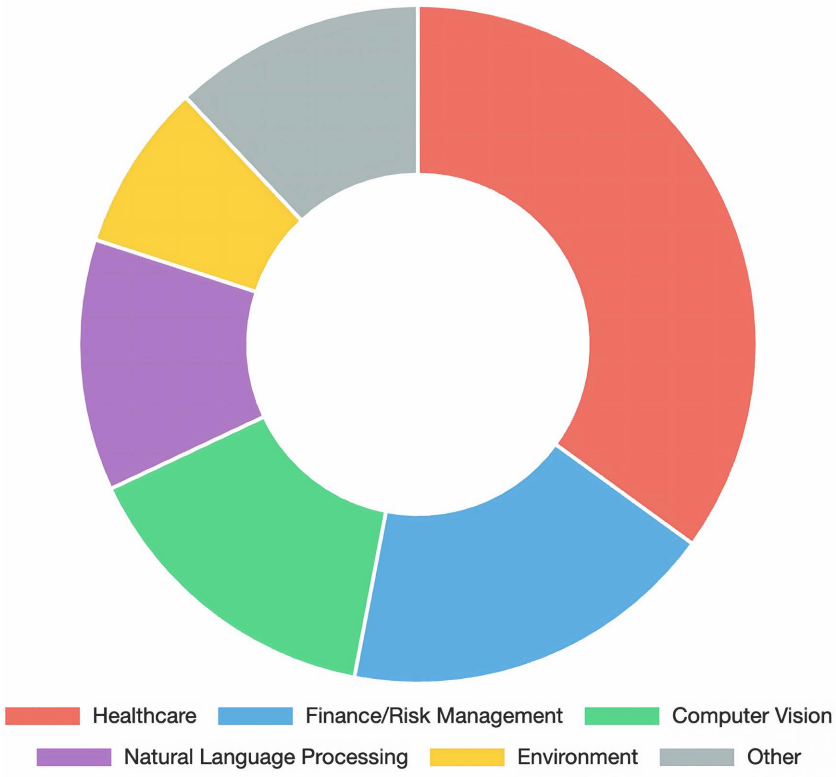

Figure 2. XAI applications across risk management domains. Data from systematic literature review by (Salih et al., 2025) analyzing 512 peer-reviewed XAI application papers, with finance/risk applications comprising 18% of total studies reviewed.

In credit scoring, XAI techniques help improve the transparency of machine learning models that predict an individual's creditworthiness (Bücker et al., 2022). Traditional credit scoring models often rely on limited data and have been criticized for being opaque in their decision-making. With XAI, lenders can use machine learning models that incorporate a wider range of data, such as spending habits and behavioral data, while still ensuring that these models provide clear, understandable reasons for their decisions. For example, SHAP values can be used to break down the factors contributing to a credit score prediction, explaining how features like income, credit history, and loan amounts influence the outcome (Moscato, Picariello, & Sperlí, 2021). This transparency not only increases trust among consumers but also helps financial institutions comply with fair lending

practices, ensuring that no discrimination takes place based on race, gender, or other inappropriate factors.

In insurance underwriting, XAI allows insurance companies to automate risk assessments while maintaining transparency in premium calculations (Maier et al., 2020). Machine learning models in underwriting are used to evaluate an applicant's risk profile and set premiums accordingly. However, if an applicant's premium is set too high or if a claim is rejected, they must understand the reasoning behind the decision. With XAI, insurers can provide explanations for their decisions, showing exactly which features, such as age, health history, or coverage type, influenced the final outcome. This not only builds customer trust but also helps meet regulatory requirements for fairness in the insurance industry.

In market risk forecasting, XAI improves the ability to explain market predictions made by machine learning models (Ohana et al., 2021). For instance, when predicting financial market trends or asset volatility, deep learning models can provide highly accurate predictions. However, financial decision-makers need to understand the reasons behind these predictions in order to make informed decisions. Explainability in these models enables stakeholders to identify the key variables that influenced the forecast, such as macroeconomic indicators, historical price trends, and market sentiment. This transparency supports better decision-making and allows for a more comprehensive understanding of potential market risks.

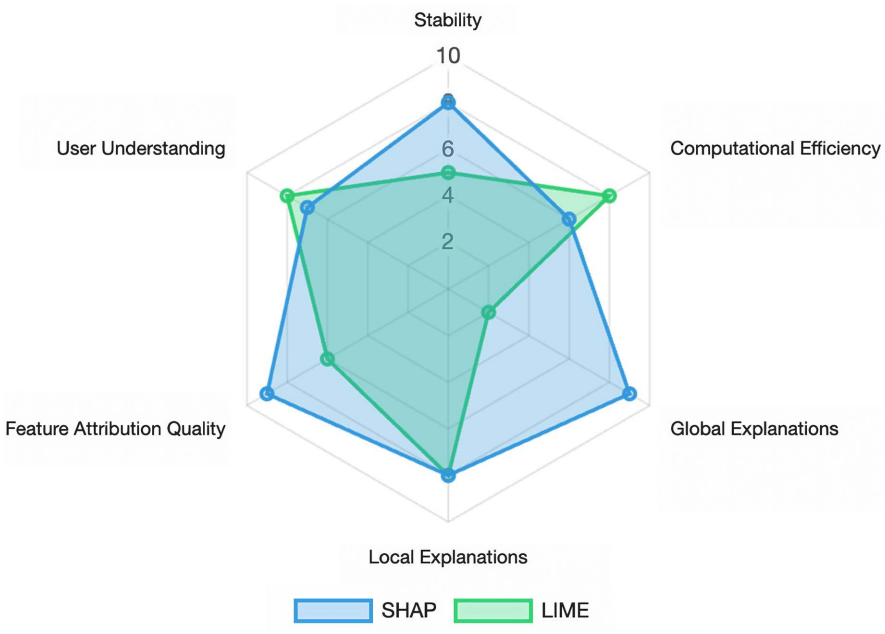

Figure 3. XAI implementation benefits in risk management. Based on empirical comparative study by (Salih et al., 2025) using biomedical dataset with 1500 subjects, and systematic evaluation framework from (Hoffman et al., 2023).

The measurement framework employed 5-point Likert scale ratings for transparency improvement, decision confidence enhancement, and regulatory compli-

ance effectiveness (Figure 3).

Overall, the application of XAI in risk management enhances decision-making by providing transparency, which is crucial in high-stakes environments. Risk managers can trust the decisions made by AI systems, knowing that they are backed by understandable and auditable processes. This trust not only boosts the confidence of stakeholders, including regulators and customers, but also ensures that the decision-making processes comply with ethical standards and legal regulations.

6. Future Directions

Research Priorities

The immediate research focus should center on developing financial services-specific evaluation metrics that go beyond current general measures, creating real-time XAI algorithms capable of sub-10-millisecond explanation generation for trading applications, and establishing formal verification frameworks that ensure XAI explanations meet diverse regulatory requirements across global jurisdictions.

Medium-term research objectives emphasize adaptive explanation systems that automatically tailor outputs based on user expertise and decision context, advancing from correlation-based to causal inference methods for deeper financial insights, and developing multi-stakeholder systems that simultaneously serve regulators, risk managers, and customers with consistent yet customized explanations.

Implementation Roadmap

Practitioners should adopt phased deployment strategies beginning with low-risk applications, establish quantitative explanation quality metrics, and invest substantially in domain-specific training programs. Technology leaders must plan for significant computational overhead (25% - 30% additional resources), implement sophisticated model monitoring with monthly recalibration schedules, and develop standardized explanation formats for system integration.

Regulatory compliance officers face the challenge of maintaining comprehensive audit trails with seven-year retention periods, establishing quarterly validation frameworks using independent datasets, and developing systems that satisfy multiple jurisdictional requirements simultaneously across US, EU, and Asia-Pacific markets.

Industry Collaboration

The future success of XAI in financial services depends on coordinated standardization efforts including industry-wide evaluation benchmarks, regulatory sandbox programs for testing innovative approaches, and data sharing consortiums that enable collaborative research while preserving competitive advantages. Technology development should prioritize open-source financial extensions to existing XAI libraries, foster specialized vendor ecosystems focused on financial requirements, and establish sustained academic partnerships for long-term research advancement.

These directions collectively aim to transform XAI from experimental technology into standard practice for transparent, accountable, and regulatory-compliant financial decision-making systems.

7. Conclusion

In conclusion, XAI plays a transformative role in risk management by improving both the accuracy and transparency of decision-making processes. While machine learning models offer significant advancements in predicting and assessing risks, the complexity of these models often makes them difficult to interpret. This lack of interpretability raises concerns, particularly in sectors like finance, insurance, and healthcare, where decisions have significant financial and regulatory implications.

By integrating XAI techniques, such as LIME and SHAP, organizations can enhance the explainability of machine learning models without sacrificing performance. This allows risk managers to gain a deeper understanding of how decisions are made, improving their ability to trust and verify the predictions made by AI systems. In high-stakes areas like fraud detection, credit scoring, and insurance underwriting, the need for transparency is critical, and XAI ensures that the decision-making process is both auditable and justifiable.

However, balancing the trade-off between accuracy and interpretability remains a challenge. While more complex models like neural networks offer superior accuracy, they often lack the transparency required for compliance and trust. On the other hand, simpler models are easier to interpret but may sacrifice the level of detail needed for accurate predictions. The future of XAI in risk management will depend on the development of explainable deep learning models and other advanced techniques that strike a better balance between performance and transparency.

Despite these challenges, XAI holds the potential to revolutionize the field of risk management by making AI models more understandable, reliable, and accountable. As organizations continue to rely on AI to manage risks, the integration of explainability into machine learning models will ensure that these systems are not only effective but also ethical, fair, and regulatory-compliant. The future of XAI in risk management is bright, with ongoing advancements likely to lead to even more transparent and trustworthy AI-driven decision-making systems.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Addy, W. A., Ajayi-Nifise, A. O., Bello, B. G., Tula, S. T., Odeyemi, O., & Falaiye, T. (2024a). AI in Credit Scoring: A Comprehensive Review of Models and Predictive Analytics. *Global Journal of Engineering and Technology Advances, 18,* 118-129. https://doi.org/10.30574/gjeta.2024.18.2.0029

Addy, W. A., Ajayi-Nifise, A. O., Bello, B. G., Tula, S. T., Odeyemi, O., & Falaiye, T.

(2024b). Machine Learning in Financial Markets: A Critical Review of Algorithmic Trading and Risk Management. *International Journal of Science and Research Archive, 11,* 1853-1862. https://doi.org/10.30574/ijsra.2024.11.1.0292

Ahmad, T., Katari, P., Pamidi Venkata, A. K., Ravi, C., & Shaik, M. (2024). Explainable AI: Interpreting Deep Learning Models for Decision Support. *Advances in Deep Learning Techniques, 4,* 80-108.

Badhon, B., Chakrabortty, R. K., Anavatti, S. G., & Vanhoucke, M. (2025). A Multi-Module Explainable Artificial Intelligence Framework for Project Risk Management: Enhancing Transparency in Decision-making. *Engineering Applications of Artificial Intelligence, 148,* Article ID: 110427. https://doi.org/10.1016/j.engappai.2025.110427

Balbaa, M. E., Astanakulov, O., Ismailova, N., & Batirova, N. (2023). Real-time Analytics in Financial Market Forecasting: A Big Data Approach. In *Proceedings of the 7th International Conference on Future Networks and Distributed Systems* (pp. 230-233). ACM. https://doi.org/10.1145/3644713.3644743

Barnes, E., & Hutson, J. (2024). Navigating the Complexities of AI: The Critical Role of Interpretability and Explainability in Ensuring Transparency and Trust. *International Journal of Multidisciplinary and Current Educational Research, 6,* 248-256.

Bello, O. A. (2023). Machine Learning Algorithms for Credit Risk Assessment: An Economic and Financial Analysis. *International Journal of Management, 10,* 109-133.

Bello, O. A., Folorunso, A., Onwuchekwa, J., Ejiofor, O. E., Budale, F. Z., & Egwuonwu, M. N. (2023). Analysing the Impact of Advanced Analytics on fraud Detection: A Machine Learning Perspective. *European Journal of Computer Science and Information Technology, 11,* 103-126.

Bhattacharya, A. (2022). *Applied Machine Learning Explainability Techniques: Make ML Models Explainable and Trustworthy for Practical Applications Using LIME, SHAP, and More.* Packt Publishing Ltd.

Borketey, B. (2024). Real-time Fraud Detection Using Machine Learning. *Journal of Data Analysis and Information Processing, 12,* 189-209. https://doi.org/10.4236/jdaip.2024.122011

Bücker, M., Szepannek, G., Gosiewska, A., & Biecek, P. (2022). Transparency, Auditability, and Explainability of Machine Learning Models in Credit Scoring. *Journal of the Operational Research Society, 73,* 70-90. https://doi.org/10.1080/01605682.2021.1922098

Burkart, N., & Huber, M. F. (2021). A Survey on the Explainability of Supervised Machine Learning. *Journal of Artificial Intelligence Research, 70,* 245-317. https://doi.org/10.1613/jair.1.12228

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics, 8,* Article 832. https://doi.org/10.3390/electronics8080832

Chamola, V., Hassija, V., Sulthana, A. R., Ghosh, D., Dhingra, D., & Sikdar, B. (2023). A Review of Trustworthy and Explainable Artificial Intelligence (xai). *IEEE Access, 11,* 78994-79015. https://doi.org/10.1109/ACCESS.2023.3294569

Dieber, J., & Kirrane, S. (2020). *Why Model Why? Assessing the Strengths and Limitations of LIME.* arXiv: 2012.00093

Dlamini, A. (2024). Machine Learning Techniques for Optimizing Recurring Billing and Revenue Collection in SaaS Payment Platforms. *Journal of Computational Intelligence, Machine Reasoning, and Decision-Making, 9,* 1-14.

Dziugaite, G. K., Ben-David, S., & Roy, D. M. (2020). *Enforcing Interpretability and Its Statistical Impacts: Trade-Offs between Accuracy and Interpretability.* arXiv: 2010.13764

Fritz-Morgenthal, S., Hein, B., & Papenbrock, J. (2022). Financial Risk Management and Explainable, Trustworthy, Responsible Ai. *Frontiers in Artificial Intelligence, 5,* Article 779799. https://doi.org/10.3389/frai.2022.779799

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K. et al. (2024). Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cognitive Computation, 16,* 45-74. https://doi.org/10.1007/s12559-023-10179-8

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2023). Measures for Explainable AI: Explanation Goodness, User Satisfaction, Mental Models, Curiosity, Trust, and Human-AI Performance. *Frontiers in Computer Science, 5,* Article 1096357. https://doi.org/10.3389/fcomp.2023.1096257

Hong, S. R., Hullman, J., & Bertini, E. (2020). Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs. *Proceedings of the ACM on Human-Computer Interaction, 4,* 1-26. https://doi.org/10.1145/3392878

Hosain, M. T., Jim, J. R., Mridha, M. F., & Kabir, M. M. (2024). Explainable AI Approaches in Deep Learning: Advancements, Applications and Challenges. *Computers and Electrical Engineering, 117,* Article ID: 109246. https://doi.org/10.1016/j.compeleceng.2024.109246

Huang, A. A., & Huang, S. Y. (2023). Increasing Transparency in Machine Learning through Bootstrap Simulation and Shapely Additive Explanations. *PLOS ONE, 18,* e0281922. https://doi.org/10.1371/journal.pone.0281922

Huang, X. (2024). Predictive Models: Regression, Decision Trees, and Clustering. *Applied and Computational Engineering, 79,* 124-133. https://doi.org/10.54254/2755-2721/79/20241551

Kapale, R., Deshpande, P., Shukla, S., Kediya, S., Pethe, Y., & Metre, S. (2024). Explainable AI for Fraud Detection: Enhancing Transparency and Trust in Financial Decision-making. In *2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI)* (pp. 1-6). IEEE. https://doi.org/10.1109/idicaiei61867.2024.10842874

Leo, M., Sharma, S., & Maddulety, K. (2019). Machine Learning in Banking Risk Management: A Literature Review. *Risks, 7,* Article 29. https://doi.org/10.3390/risks7010029

Lisboa, P. J. G., Saralajew, S., Vellido, A., Fernández-Domenech, R., & Villmann, T. (2023). The Coming of Age of Interpretable and Explainable Machine Learning Models. *Neurocomputing, 535,* 25-39. https://doi.org/10.1016/j.neucom.2023.02.040

Maier, M., Carlotto, H., Saperstein, S., Sanchez, F., Balogun, S., & Merritt, S. (2020). Improving the Accuracy and Transparency of Underwriting with Artificial Intelligence to Transform the Life-insurance Industry. *AI Magazine, 41,* 78-93. https://doi.org/10.1609/aimag.v41i3.5320

Malhotra, R., & Malhotra, D. K. (2023). The Impact of Technology, Big Data, and Analytics: The Evolving Data-Driven Model of Innovation in the Finance Industry. *The Journal of Financial Data Science, 5,* 50-65. https://doi.org/10.3905/jfds.2023.1.129

Moscato, V., Picariello, A., & Sperlí, G. (2021). A Benchmark of Machine Learning Approaches for Credit Score Prediction. *Expert Systems with Applications, 165,* Article ID: 113986. https://doi.org/10.1016/j.eswa.2020.113986

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y. et al. (2023). From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Computing Surveys, 55,* 1-42. https://doi.org/10.1145/3583558

Odonkor, B., Kaggwa, S., Uwaoma, P. U., Hassan, A. O., & Farayola, O. A. (2024). The Impact of AI on Accounting Practices: A Review: Exploring How Artificial Intelligence Is Transforming Traditional Accounting Methods and Financial Reporting. *World Journal of Advanced Research and Reviews, 21,* 172-188. https://doi.org/10.30574/wjarr.2024.21.1.2721

Oguntibeju, O. O. (2024). Mitigating Artificial Intelligence Bias in Financial Systems: A Comparative Analysis of Debiasing Techniques. *Asian Journal of Research in Computer Science, 17,* 165-178. https://doi.org/10.9734/ajrcos/2024/v17i12536

Ohana, J. J., Ohana, S., Benhamou, E., Saltiel, D., & Guez, B. (2021). Explainable AI (XAI) Models Applied to the Multi-Agent Environment of Financial Markets. In D. Calvaresi, A. Najjar, M. Winikoff, & K. Främling (eds.), *Explainable and Transparent AI and Multi-Agent Systems. EXTRAAMAS 2021* (pp. 189-207). Springer International Publishing. https://doi.org/10.1007/978-3-030-82017-6_12

Olushola, A., & Mart, J. (2024). *Fraud Detection using Machine Learning*. Science Open.

Rane, N., Choudhary, S., & Rane, J. (2023). Explainable Artificial Intelligence (XAI) Approaches for Transparency and Accountability in Financial Decision-Making. *SSRN Electronic Journal*. (Preprint) https://doi.org/10.2139/ssrn.4640316

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. https://doi.org/10.1038/s42256-019-0048-x

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges. *Statistics Surveys, 16,* 1-85. https://doi.org/10.1214/21-ss133

Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K. et al. (2025). A Perspective on Explainable Artificial Intelligence Methods: SHAP and Lime. *Advanced Intelligent Systems, 7,* Article ID: 2400304. https://doi.org/10.1002/aisy.202400304

Vimbi, V., Shaffi, N., & Mahmud, M. (2024). Interpreting Artificial Intelligence Models: A Systematic Review on the Application of LIME and SHAP in Alzheimer's Disease Detection. *Brain Informatics, 11,* Article No. 10. https://doi.org/10.1186/s40708-024-00222-1

Wilson, A., & Anwar, M. R. (2024). The Future of Adaptive Machine Learning Algorithms in High-Dimensional Data Processing. *International Transactions on Artificial Intelligence (ITALIC), 3,* 97-107. https://doi.org/10.33050/italic.v3i1.656