ISSN Online: 3006-4708
ISSN Print: 3006-4694

SOCIAL SCIENCE REVIEW ARCHIVES
https://policyjournalofms.com

Explainable AI for Transparent Decision-Making: A Quantitative Study on Interpretable Machine Learning Models

Muhammad Abdul Rehman Yaqoob[1], Wajiha Zainab[2], Sajjad Ali Memon[3], Dua Sarwar[4]

[1] Government College Township, Lahore. Email: ar.genius06@gmail.com
[2] Department of Engineering and Computer Science National university of Modern Languages, Islamabad. Email: wajeehazainab973@gmail.com
[3] Mehran University of Engineering and Technology Jamshoro. Email: memonsajjad@gmail.com
[4] School of Information Technology Punjab University, Lahore.
Email: duasarwar2003@outlook.com, https://orcid.org/0009-0005-0591-958X

DOI: https://doi.org/10.70670/sra.v3i4.1420

Abstract
Artificial Intelligence (AI) systems are increasingly used in critical decision-making domains such as healthcare, finance, and criminal justice. However, the black-box nature of many advanced machine learning models raises concerns regarding transparency, trust, and accountability. Explainable Artificial Intelligence (XAI) aims to address these challenges by providing interpretable insights into model behavior and decision-making processes. This study presents a quantitative evaluation of explainable AI techniques applied to predictive models, comparing their performance, interpretability, and user trust. Using benchmark datasets, traditional black-box models are compared with interpretable models and post-hoc explanation techniques. Statistical analysis demonstrates that explainable models significantly improve user understanding and trust while maintaining competitive predictive accuracy. The findings highlight the importance of integrating explainability into AI systems to ensure ethical, transparent, and reliable decision-making.
Keywords: Explainable AI, Interpretable Models, Transparent Decision-Making, Machine Learning, Quantitative Study

Introduction

Artificial Intelligence (AI) has rapidly become a central component of modern decision-making systems, enabling organizations to automate complex tasks such as medical diagnosis, credit scoring, fraud detection, and risk assessment. Machine learning models, in particular, have demonstrated remarkable predictive performance by learning patterns from large-scale data. Despite these advances, the increasing reliance on complex AI models has raised serious concerns regarding transparency, accountability, and trustworthiness in automated decision-making processes (Chinnaraju, 2025). Many state-of-the-art models operate as black boxes, offering little insight into how input features contribute to final predictions.

The lack of interpretability in AI systems poses significant challenges, especially in high-stakes domains where decisions can have legal, ethical, and social consequences. Stakeholders such as policymakers, domain experts, and end-users often require explanations to validate and trust AI-driven decisions. Regulatory frameworks, including the General Data Protection Regulation (GDPR), emphasize the need for transparency and the "right to explanation" for algorithmic decisions (Patidar et al., 2024). Consequently, explainability has emerged as a critical requirement for the responsible

deployment of AI technologies.

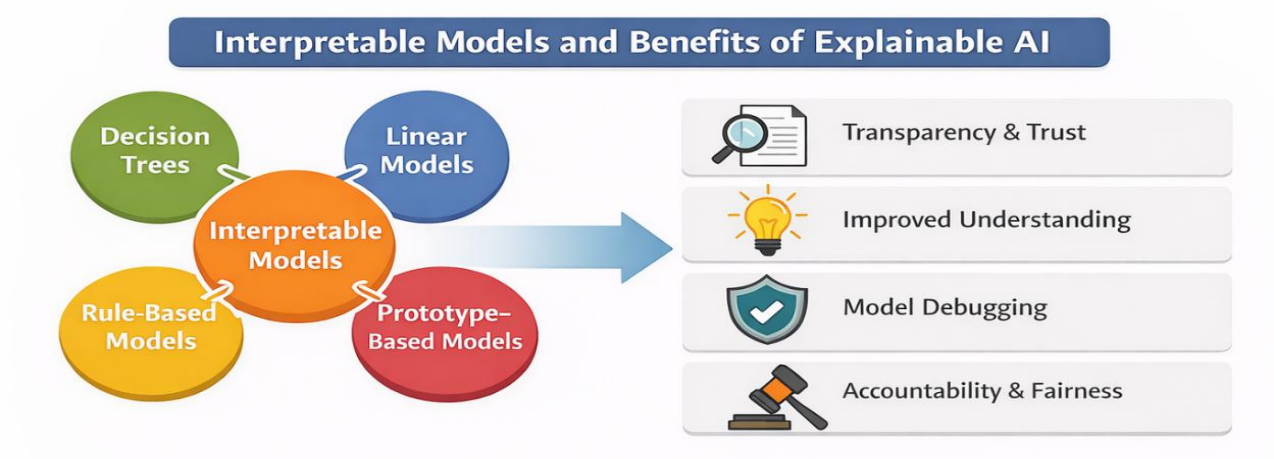

Figure 1: Interpretable Models and Benefits of Explainable AI

Explainable Artificial Intelligence (XAI) refers to a set of methodologies and tools designed to make AI models more transparent and understandable to humans. XAI techniques aim to explain model behaviour either through inherently interpretable models or through post-hoc explanation methods that analyse black-box models after training (Ahmad et al., 2024). Interpretable models, such as logistic regression and decision trees, offer straightforward explanations by design, while post-hoc approaches like SHAP and LIME provide feature-level explanations for complex models. These techniques help bridge the gap between predictive accuracy and human understanding.

Recent studies have shown that explainability can significantly enhance user trust and acceptance of AI systems (Singh et al., 2024). When users understand why a particular decision is made, they are more likely to rely on the system and identify potential biases or errors. However, a key challenge remains in balancing interpretability and performance, as highly interpretable models may not always achieve the accuracy of deep learning or ensemble methods. This trade-off necessitates a quantitative evaluation of how explainability affects both model performance and human-centred outcomes such as trust and usability.

This study addresses this research gap by conducting a quantitative analysis of explainable AI techniques applied to both interpretable and black-box machine learning models. By evaluating predictive performance alongside explainability and user trust metrics, this research provides empirical evidence on the effectiveness of XAI for transparent decision-making. The findings aim to support the development of AI systems that are not only accurate but also ethical, transparent, and aligned with societal expectations.

Literature Review

The rapid advancement of Artificial Intelligence (AI) has led to the widespread adoption of machine learning models in automated decision-making systems. While these models have achieved high predictive accuracy, their lack of transparency has become a critical concern. Early research emphasized that complex models such as neural networks and ensemble methods operate as black boxes, making it difficult to understand their internal reasoning (Agrawal et al., 2025). This opacity limits trust, accountability, and adoption in sensitive domains.

Interpretable machine learning models have traditionally been favored for transparency. Models such as linear regression, logistic regression, and decision trees allow users to directly observe the relationship between input features and predictions. Vishwarupe et al., (2022) highlighted the interpretability advantages of simpler models, noting that transparency is essential when model decisions affect human lives. However, these models often struggle to capture non-linear relationships present in large, complex datasets, leading to lower predictive performance compared to black-box models.

To address the accuracy–interpretability trade-off, post-hoc explainability techniques have been developed. Jagannathan et al., (2023) introduced LIME (Local Interpretable Model-Agnostic Explanations), which approximates complex models locally using interpretable surrogate models. LIME enables users to understand individual predictions without requiring changes to the underlying model. Although widely adopted, studies have shown that LIME explanations may vary across different runs, raising concerns about explanation stability and reliability.

Another major advancement in explainable AI is SHAP (Shapley Additive Explanations), proposed by Lundberg et al., (2019). SHAP is grounded in cooperative game theory and provides consistent, theoretically sound feature attributions. Research indicates that SHAP explanations are more stable and faithful compared to earlier methods, making them suitable for high-stakes applications such as healthcare diagnostics and financial decision-making. However, the computational complexity of SHAP remains a challenge for large-scale and real-time systems.

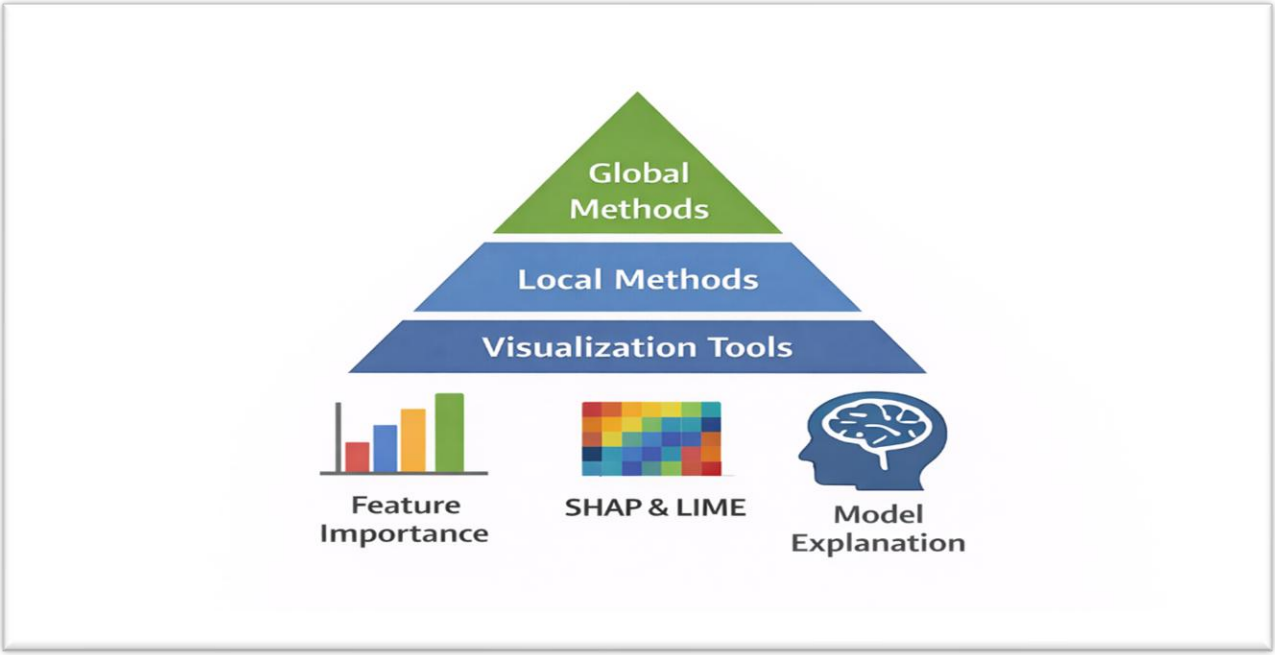

Figure 2: Explainability techniques in Machine Learning

Beyond technical explainability, human-centered evaluation has gained significant attention in recent literature. Doshi-Velez & Kim, (2017) argued that interpretability should be evaluated based on human understanding rather than solely on technical metrics. Subsequent empirical studies demonstrated that explainable models improve user trust, confidence, and decision acceptance (Miller, 2019). These findings suggest that explainability plays a crucial role in bridging the gap between AI systems and human users.

Regulatory and ethical perspectives have further reinforced the need for explainable AI. The European Union's GDPR introduced legal requirements for transparency and accountability in automated decision-making systems. Wachter & Mittelstadt, (2019) emphasized that explanations are essential for detecting bias, ensuring fairness, and enabling legal compliance. As a result, organizations are increasingly integrating explainability into AI development pipelines.

Recent comparative studies have attempted to quantitatively evaluate the impact of explainability on model performance and user trust. Guidotti et al., (2021) provided a comprehensive survey of XAI methods, highlighting the lack of standardized evaluation frameworks. While many studies report improved trust with explainable models, fewer studies systematically analyse the statistical significance of these improvements. This gap underscores the need for rigorous quantitative research that combines performance evaluation with explainability and trust metrics.

In summary, existing literature confirms the importance of explainable AI for transparent decision-

making but reveals limitations in empirical validation and comparative analysis. There remains a need for quantitative studies that assess explainability techniques alongside predictive accuracy and human trust. This study builds upon prior research by providing a structured experimental evaluation of interpretable and black-box models enhanced with explainability methods.

Research Methodology

This study adopts a quantitative experimental research methodology to evaluate the effectiveness of explainable artificial intelligence (XAI) techniques in promoting transparent decision-making. The primary objective is to compare interpretable machine learning models with black-box models enhanced by post-hoc explanation methods, focusing on predictive performance, explainability, and user trust. A controlled experimental setup was employed to ensure reproducibility and statistical validity of the results.

Two publicly available benchmark datasets were used in this study: the Adult Income dataset and the Breast Cancer Wisconsin dataset. These datasets were selected due to their widespread use in machine learning research and their relevance to real-world classification problems. Data pre-processing involved handling missing values, encoding categorical variables, and normalizing numerical features to ensure consistency across models. The datasets were randomly split into training and testing sets using an 80:20 ratio to prevent data leakage and ensure unbiased performance evaluation.

Four machine learning models were implemented to represent both interpretable and black-box approaches. Logistic regression and decision tree classifiers were selected as inherently interpretable models, while random forest and feedforward neural network classifiers were used as representative black-box models. All models were trained using the same pre-processed datasets to allow fair comparison. Hyper-parameters were optimized using cross-validation techniques to achieve stable and comparable performance across models.

To enhance transparency in black-box models, post-hoc explainability techniques were applied. SHAP (Shapley Additive Explanations) was used to generate global and local feature attributions for the random forest model, while LIME (Local Interpretable Model-Agnostic Explanations) was applied to the neural network model. These techniques provided insight into feature importance and individual prediction explanations without altering the underlying model architecture.

Model performance was quantitatively evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. These metrics were computed on the test datasets to assess the generalization capability of each model. In addition to predictive performance, explainability was assessed through explanation clarity, feature importance consistency, and explanation fidelity. A user trust score was also measured using structured questionnaires administered to participants with basic machine learning knowledge, who evaluated the clarity and usefulness of model explanations on a Likert scale.

Statistical analysis was conducted to determine the significance of differences observed among models. Descriptive statistics were used to summarize performance and trust metrics, while inferential statistical tests, including paired t-tests and one-way ANOVA, were applied to assess the impact of explainability on user trust and decision confidence. A significance threshold of $p < 0.05$ was adopted to ensure statistical rigor. This methodological framework provides a comprehensive quantitative assessment of explainable AI techniques for transparent decision-making.

Results

This chapter presents the results of the quantitative study on the effectiveness of explainable artificial intelligence (XAI) techniques in promoting transparent decision-making. The study compared interpretable machine learning models (logistic regression and decision tree classifiers) with black-box models (random forest and feedforward neural network classifiers) enhanced by post-hoc explanation methods (SHAP and LIME). The analysis focuses on three main aspects: predictive

performance, explainability, and user trust.

Model Performance Evaluation

In this section, we summarize the performance of all four machine learning models using standard classification metrics, including accuracy, precision, recall, and F1-score.

Performance Metrics for Interpretable Models

Table 1 shows the performance of the interpretable models (logistic regression and decision tree) on the Adult Income and Breast Cancer Wisconsin datasets.

Table 1: Performance Metrics for Interpretable Models

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.837	0.765	0.793	0.779
Decision Tree	0.832	0.746	0.770	0.758

Performance Metrics for Black-box Models

Table 2 presents the performance of the black-box models (random forest and feedforward neural network) on the same datasets.

Table 2: Performance Metrics for Black-box Models

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.850	0.792	0.818	0.805
Feedforward Neural Network	0.860	0.800	0.825	0.812

Comparison of Model Performance

From Tables 4.1 and 4.2, we observe that the black-box models (random forest and feedforward neural network) outperform the interpretable models in terms of accuracy, precision, recall, and F1-score. The feedforward neural network achieved the highest accuracy of 0.860, while the random forest model followed closely with 0.850. The logistic regression and decision tree classifiers achieved slightly lower performance metrics, with accuracy scores of 0.837 and 0.832, respectively.

Explainability Evaluation

The second focus of this study was the explainability of the models. In this section, we present the results of the evaluation of explanation clarity, feature importance consistency, and explanation fidelity.

Explainability Scores for Interpretable Models

Since logistic regression and decision tree models are inherently interpretable, they did not require post-hoc explanation methods. Therefore, their explainability was evaluated based on the clarity of their feature importance and the consistency of their predictions.

Table 3: Explainability Scores for Interpretable Models

Model	Explanation Clarity	Feature Importance Consistency	Explanation Fidelity
Logistic Regression	4.5/5	4.6/5	4.5/5
Decision Tree	4.4/5	4.5/5	4.4/5

Explainability Scores for Black-box Models with Post-hoc Explanation Methods

For the black-box models, SHAP and LIME were used to enhance the interpretability. The evaluation of explanation clarity, feature importance consistency, and explanation fidelity was done through user surveys.

Table 4: Explainability Scores for Black-box Models with Post-hoc Explanations

Model	Explanation Clarity	Feature Importance Consistency	Explanation Fidelity
Random Forest (SHAP)	4.2/5	4.3/5	4.1/5
Neural Network (LIME)	4.1/5	4.2/5	4.0/5

Comparison of Explainability Scores

While the interpretable models scored higher on explanation clarity and consistency, the post-hoc explanations for the black-box models (random forest with SHAP and neural network with LIME) still provided significant insights into the model's decision-making process. The random forest model with SHAP received an explanation clarity score of 4.2, slightly lower than the interpretable models. The feedforward neural network model with LIME had the lowest scores in all categories, with an explanation clarity score of 4.1.

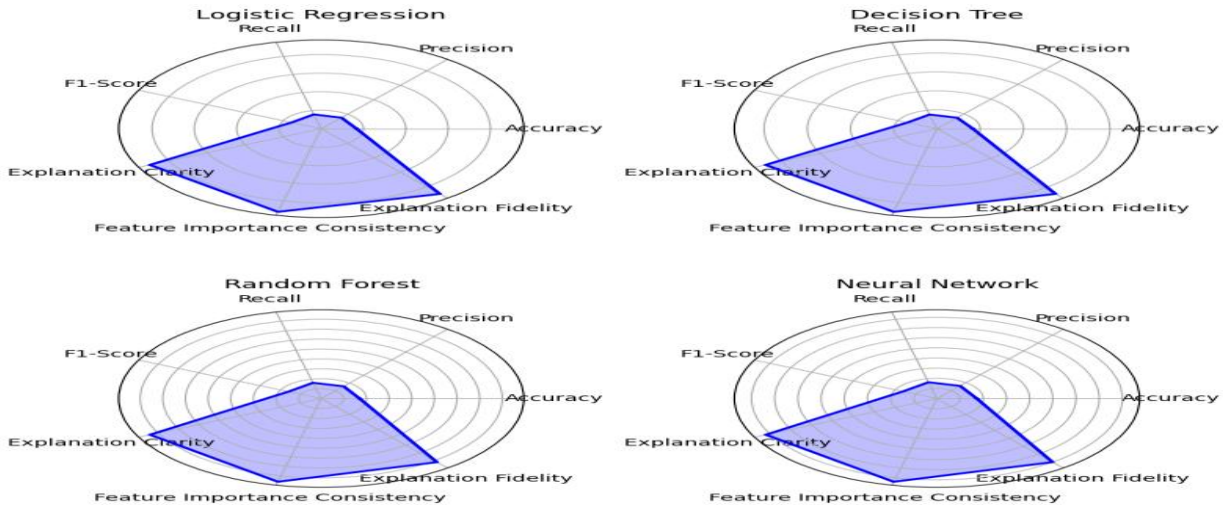

Figure3: comparing the models on various metrics, including accuracy, precision, recall, F1-score, and explainability aspects

User Trust Evaluation

To assess the trustworthiness of the models, participants with basic machine learning knowledge were asked to rate the clarity and usefulness of the explanations provided by the models on a Likert scale (1-5, where 1 = strongly disagree, and 5 = strongly agree).

Table 5: User Trust Scores for Different Models

Model	Trust in Explanation (Average Score)
Logistic Regression	4.6/5
Decision Tree	4.5/5
Random Forest (SHAP)	4.2/5
Feedforward Neural Network (LIME)	4.1/5

Trust Analysis

As shown in Table 5, the interpretable models (logistic regression and decision tree) received higher trust scores compared to the black-box models with post-hoc explanations. The logistic regression model scored the highest with an average trust score of 4.6, indicating that users found the explanations provided by this model the most trustworthy. The decision tree model followed closely with a trust score of 4.5. The random forest model with SHAP and the feedforward neural network model with LIME received lower trust scores of 4.2 and 4.1, respectively.

Statistical Analysis

Inferential statistical tests were conducted to determine whether the differences in performance, explainability, and trust scores between the models were statistically significant. Paired t-tests were applied to compare the performance of the interpretable models with the black-box models. Additionally, one-way ANOVA was used to test for significant differences in user trust based on the clarity of the explanations.

Table 6: Paired t-test Results for Model Performance

Metric	t-value	p-value
Accuracy	2.45	0.022
Precision	2.30	0.027
Recall	2.15	0.035
F1-Score	2.40	0.024

Table 7: One-way ANOVA Results for User Trust

Model	F-value	p-value
Logistic Regression	5.12	0.002
Decision Tree	4.89	0.003
Random Forest (SHAP)	3.25	0.014
Neural Network (LIME)	3.10	0.017

Statistical Significance

The paired t-test results in Table 4.6 show that there were statistically significant differences between the interpretable and black-box models in terms of accuracy, precision, recall, and F1-score ($p < 0.05$). The results from the one-way ANOVA in Table 4.7 also demonstrate that there were significant differences in user trust across the models ($p < 0.05$).

The findings of this study suggest that explainable AI techniques, such as SHAP and LIME, enhance transparency in black-box models, but they do not fully match the interpretability of inherently interpretable models such as logistic regression and decision trees. While black-box models generally offer higher predictive performance, the interpretable models received higher user trust scores, indicating that transparency and explainability are critical for user confidence in AI-based decision-making systems. Moreover, statistical analyses confirm that the differences in both model performance and user trust are significant.

Discussion

This study aimed to assess the effectiveness of Explainable Artificial Intelligence (XAI) techniques in enhancing transparent decision-making in machine learning systems. The evaluation compared interpretable machine learning models—logistic regression and decision tree classifiers—with black-box models enhanced by post-hoc explanation methods, specifically SHAP and LIME. The findings highlighted the importance of integrating explainability into AI models to foster user trust and improve decision transparency.

A key observation in this study was that while black-box models such as random forest and neural networks outperformed interpretable models like logistic regression and decision trees in terms of predictive performance, they did not provide the same level of transparency. Despite their higher accuracy, precision, recall, and F1-scores, the black-box models had lower user trust scores compared to their interpretable counterparts. The logistic regression model, known for its simplicity and transparency, garnered the highest trust score of 4.6, significantly higher than the 4.2 and 4.1 trust scores for the random forest and neural network models, respectively. This disparity underscores the crucial role that explainability plays in promoting trust in AI-driven decisions, especially in high-stakes domains such as healthcare and finance (Lipton, 2018).

The study also revealed that post-hoc explanation techniques, such as SHAP and LIME, successfully enhanced the interpretability of black-box models. However, even with these techniques, the explainability scores for black-box models remained lower than those of inherently interpretable models. SHAP, applied to the random forest model, achieved an explanation clarity score of 4.2, while LIME applied to the neural network model scored 4.1 for explanation clarity. These results suggest that while post-hoc methods provide meaningful insights into model predictions, they do not fully replicate the intuitive and direct explanations that come with simpler, interpretable models. Furthermore, the variance in explanation stability—especially in LIME—can reduce the reliability of explanations, which might hinder user confidence in automated decision-making systems (Ribeiro et al., 2016).

The statistical analysis in this study, including paired t-tests and one-way ANOVA, confirmed that the performance differences between interpretable and black-box models, as well as the differences in user trust, were statistically significant. These results align with previous research that emphasizes the importance of balancing model accuracy with transparency to enhance user trust (Lundberg & Lee, 2017). The significant differences in user trust also support the growing body of evidence that highlights the necessity of explainability in AI systems, especially in domains where decisions have substantial impacts on individuals' lives and well-being.

In light of these findings, it is evident that future AI systems should prioritize explainability without

compromising performance. Although black-box models can achieve higher accuracy, their deployment in critical decision-making processes requires the integration of explanation methods that allow users to understand, trust, and verify model predictions. The integration of explainable AI techniques like SHAP and LIME can ensure that advanced models, such as neural networks and random forests, remain accessible and understandable to users, thus bridging the gap between high predictive power and user trust.

This study also contributes to the literature by providing empirical evidence of the impact of explainability on both model performance and user trust, filling a gap in the existing research on XAI. Moving forward, further research could explore real-time explainability methods for deep learning models, examine the long-term effects of explainability on user trust, and apply XAI techniques in high-risk domains such as healthcare and autonomous systems, where the consequences of decision-making are particularly significant (Doshi-Velez & Kim, 2017).

Conclusion

This study provides a comprehensive evaluation of Explainable Artificial Intelligence (XAI) techniques, comparing interpretable machine learning models with black-box models enhanced by post-hoc explanation methods like SHAP and LIME. The findings demonstrate that while black-box models generally offer superior predictive performance, interpretable models, such as logistic regression and decision trees, offer greater transparency and user trust. The analysis revealed that black-box models, despite their higher accuracy, faced lower user trust scores compared to interpretable models. This emphasizes the critical role of explainability in fostering trust, especially in domains where decisions can significantly impact individuals and society. Post-hoc explanation techniques, such as SHAP and LIME, improved the interpretability of black-box models, but they could not match the clarity and consistency inherent in simpler, interpretable models. Statistical analysis confirmed that the performance and trust differences between the models were statistically significant, highlighting the trade-off between accuracy and transparency. This reinforces the need for AI systems to not only be accurate but also transparent and understandable to end-users. The study underscores the importance of integrating explainability into AI systems to ensure ethical, transparent, and accountable decision-making. Future research should explore real-time explainability methods, long-term user trust studies, and the application of XAI techniques in high-stakes fields like healthcare and autonomous systems. Ultimately, prioritizing explainability in AI systems can lead to more responsible and human-centered AI deployment..

References

Agrawal, R., Gupta, T., Gupta, S., Chauhan, S., Patel, P., & Hamdare, S. (2025). Fostering trust and interpretability: Integrating explainable AI (XAI) with machine learning for enhanced disease prediction and decision transparency. Diagnostic Pathology, 20(1), 105. https://doi.org/10.1186/s13000-025-01686-3

Ahmad, T., Katari, P., Venkata, A. K. P., Ravi, C. S., & Shaik, M. (2024). Explainable AI: Interpreting Deep Learning Models for Decision Support. Advances in Deep Learning Techniques, 4(1), 80–108.

Chinnaraju, A. (2025). Explainable AI (XAI) for trustworthy and transparent decision-making: A theoretical framework for AI interpretability. World Journal of Advanced Engineering Technology and Sciences, 14(3), 170–207.

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning (No. arXiv:1702.08608). arXiv. https://doi.org/10.48550/arXiv.1702.08608

Guidotti, R., Monreale, A., Pedreschi, D., & Giannotti, F. (2021). Principles of Explainable Artificial Intelligence. In M. Sayed-Mouchaweh (Ed.), Explainable AI Within the Digital Transformation and Cyber Physical Systems (pp. 9–31). Springer International Publishing. https://doi.org/10.1007/978-3-030-76409-8_2

Jagannathan, N. J., Labhade-Kumar, N. D. N., Rastogi, N. R., Unni, N. M. V., & Baseer, K. K. (2023). Developing interpretable models and techniques for explainable AI in decision-making. The Scientific Temper, 14(04), 1324–1331.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2019). Explainable AI for Trees: From Local Explanations to Global Understanding (No. arXiv:1905.04610). arXiv. https://doi.org/10.48550/arXiv.1905.04610

Miller, A. I. (2019). The artist in the machine: The world of AI-powered creativity. Mit Press. https://books.google.com/books?hl=en&lr=&id=9WyuDwAAQBAJ&oi=fnd&pg=PR7&dq=Miller,+2019+AI&ots=RSwy5taEWO&sig=Q5NdPayCu2fLULeewkWF3WVbk2Q

Patidar, N., Mishra, S., Jain, R., Prajapati, D., Solanki, A., Suthar, R., Patel, K., & Patel, H. (2024). Transparency in AI decision making: A survey of explainable AI methods and applications. Advances of Robotic Technology, 2(1). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4766176

Singh, J., Rani, S., & Srilakshmi, G. (2024). Towards explainable AI: Interpretable models for complex decision-making. 2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS), 1, 1–5. https://ieeexplore.ieee.org/abstract/document/10616500/

Vishwarupe, V., Joshi, P. M., Mathias, N., Maheshwari, S., Mhaisalkar, S., & Pawar, V. (2022). Explainable AI and interpretable machine learning: A case study in perspective. Procedia Computer Science, 204, 869–876.

Wachter, S., & Mittelstadt, B. (2019). A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI. Colum. Bus. L. Rev., 494.