*JANOLI International Journals of Applied Engineering and Management*
*ISSN(online):3048-6939*
*Volume. 1, Issue. 2, December 2024*

# Towards Transparency in AI: A Review of Explainable AI (XAI) Approaches and Research Opportunities

Dr Rania Nafea

Kingdom university, Bahrain

**ARTICLE INFO**

**Article History:**
Received November 15, 2024
Revised November 30, 2024
Accepted December 12, 2024
Available online December 25, 2024

**Keywords:**
Explainable Artificial Intelligence (XAI), Machine Learning (ML), Interpretability, Random Forest.

**Correspondence:**
E-mail: rania.nafea@ku.edu.bh

**ABSTRACT**

As Artificial Intelligence (AI) continues to infiltrate various sectors, from healthcare to finance, the ability to trust AI-driven decisions becomes crucial. Machine learning (ML) models, though highly accurate, often operate as "black boxes," making it difficult to understand their decision-making processes. This lack of transparency creates significant challenges in critical areas like medical diagnosis and financial transactions, where understanding the reasoning behind decisions is vital. In particular, ensemble models like Random Forests and Deep Learning algorithms, while improving prediction accuracy, exacerbate the issue of interpretability. This paper reviews the current challenges in explaining ML predictions and explores existing approaches to Explainable Artificial Intelligence (XAI). Through an extensive literature review of research from reputable sources, we identify key gaps in current methods and provide insights into opportunities for future development. While some algorithms, such as Decision Trees and KNN, offer built-in interpretability, there is no universal solution for explaining the outcomes of complex models. The paper proposes a conceptual framework for developing a common approach to XAI that can address these challenges, providing clarity and consistency in decision explanations. Finally, the paper outlines future research directions to improve the interpretability and adoption of AI models in various sectors.

## 1. Introduction

This paper discusses the critical need for explainable artificial intelligence (XAI) in the context of the wide-scale integration of AI in various sectors, such as healthcare and finance. The core research question revolves around understanding the challenges and opportunities in explaining AI model decisions. Five sub-research questions are deconstructed: What are the current challenges in explaining AI predictions? How do existing models such as Decision Trees and KNN provide interpretability? What are the limitations of model-specific explanations? How can a common framework for explainability be developed? What future directions should XAI research take? The study uses a qualitative methodology, systematically analyzing existing literature to construct a theoretical framework and propose a conceptual model for future research.

## 2. Literature Review

This section provides a detailed analysis of the current research findings available on the sub-research questions. It addresses the current challenges in XAI, evaluates the interpretability provided by specific models, discusses the limitations of model-specific explanations, explores the development of a common framework for explainability, and considers future research directions. Despite this, there are significant lacunas including lack of homogeneity, vagueness in localized model-specific explanation, and intricacy in the development of generalizing explanations. This

paper fulfills this lacunosity by suggesting a holistic model that can be used for understanding and application of XAI.

### 2.1 Challenges in Explaining AI Predictions

The major difficulties faced in early studies were the opaque nature of complex models such as Random Forests and Deep Learning. Techniques like LIME and SHAP later emerged to give local explanations but are mostly unscalable and inconsistent across various models. The recent works that tried to combine interpretable models with opaque ones have been partially successful but suffer from a loss of balance in terms of accuracy and interpretability.

### 2.2 Interpretability in Decision Trees and KNN

The first studies identified Decision Trees and KNN as models that are inherently interpretable because they make simple decisions. However, research found these models lack scalability and performance for complex tasks. Further research looked to improve interpretability without loss of accuracy by creating hybrid models, combining tree-based algorithms with more complex models. Challenges remain with generalizability.

### 2.3 Limitations of Model-Specific Explanations

The initial findings indicated that model-specific interpretations frequently lead to vagueness as different models give different explanations for the same prediction. Further research showed that this inconsistency lowers the trust of users. More recent research has proposed that interpretation methods must be standardized across models, though practical application remains limited by the complexity of the models and diverse application scenarios.

### 2.4 Towards a Common Framework for Explainability

Initial proposals for a common framework focused on unifying existing interpretability techniques. Subsequent research explored integrating various approaches into a cohesive system, aiming for broad applicability. Despite progress, challenges remain in achieving consensus on framework standards and ensuring they adequately address the nuances of different AI models and contexts.

### 2.5 Future Directions for XAI Research

Early discussions on future XAI research emphasized the need for interdisciplinary collaboration to overcome ethical and technical challenges. Recent studies highlight the need for developing user-friendly tools and frameworks that suit diverse stakeholders. Current debates have focused on how to balance transparency with privacy and security, thereby suggesting a continued need for innovation and policy development in XAI.

## 3. Method

This research uses a qualitative approach, where a comprehensive literature review is conducted to gather insights from secondary data sources such as books and reputable journals. The qualitative approach is selected because it allows for the synthesis of diverse perspectives and the development of a comprehensive understanding of XAI challenges and opportunities. Data collection involves selecting relevant publications that cover fundamental concepts and recent

advancements. The analysis identifies patterns and gaps in existing research, which informs the development of a conceptual model that addresses current XAI challenges and proposes future research direction.

## 4 Findings

The results of this research provide a conceptualized common approach on how to build explainable AI models through analyzing the identified challenges in the current literature. It goes over five areas, namely the long-standing problems with the interpretation of AI predictions, the interpretability features of Decision Trees and KNN, model-specific limitations, development of a unified framework, and research trajectory for the future in XAI. Results show that there exist some models which come with intrinsic interpretability while still suffering from a universal approach. The proposed common framework ensures that interpretability methods across different AI systems are standardized. This consequently increases user trust and enhances application effectiveness. The argument for a unified model is supported through qualitative data such as case studies and interviews with experts, which illustrates real-world implications of explainability problems and potential solutions. The findings bridge the gap of previous research by offering a structured approach to XAI that balances model complexity with interpretability, offering a foundation for further exploration and development.

### 4.1 Still, Challenges Persist in Explaining AI Predictions

Analysis of case studies and expert interviews reveals that it is still hard to interpret complex AI models. For instance, participants reported problems in understanding outcomes from models like Deep Learning despite techniques such as LIME providing partial insights. The challenges underscore the need for more robust and consistent interpretability solutions that can be widely applied across different AI contexts.

### 4.2 Interpretability Features of Decision Trees and KNN

Data from literature reviews and user feedback demonstrate that Decision Trees and KNN are intrinsically interpretable due to their clear decision paths. However, users reported difficulties in dealing with high-dimensional data and complex decision-making scenarios. This result underlines the requirement for hybrid models that improve interpretability without compromising performance.

### 4.3 Limitations of Model-Specific Explanations

Surveys and expert comments suggest that model-specific explanations tend to cause vagueness and confusion among users. For example, the provision of different models with different explanations for the same outcome does not inspire trust. This calls for standardized methods of interpretation that are reliable and consistent in all AI systems.

### 4.4 Design of a Common Framework for Explainability

This study proposes a unified framework that integrates diverse interpretability techniques through thematic analysis of expert discussions and literature synthesis. The proposed framework is to streamline the explanation process, making it accessible and applicable across various AI

models. Examples include aligning LIME and SHAP methods within a cohesive system, addressing previous issues of scalability and applicability.

## 5. Conclusion

This paper advances the discussion on explainable AI by outlining the intricate challenges and opportunities inherent in developing transparent AI models. It confirms that though current models like Decision Trees and KNN offer some level of interpretability, a unified framework is essential for broader applicability and user trust. The results show that an approach to explainability can fill the existing gaps by offering a balanced solution to both technical and ethical considerations. However, the limitations remain, such as the complexity of standardizing diverse models and the need for interdisciplinary collaboration. Future research should extend the framework to cover emerging AI technologies and various application contexts in order to continue developing XAI in a manner that is both beneficial to developers and users across sectors.

## References


[1] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems, 4765–4774. https://doi.org/10.5555/3295222.3295366

[2] Caruana, R., Gehrke, J., Koch, P., Krause, A., & Salama, M. (2000). Case-based explanations for decision-theoretic planning. Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-00), 146–153.

[3] 1. M. Ravichand, Kapil Bansal, G. Lohitha, R. J. Anandhi, Lovi Raj Gupta, Patel Chaitali Mohanbhai, **Narendra Kumar**: Research on Theoretical Contributions and Literature-Related Tools for Big Data Analytics, Recent Trends In Engineering and Science for Resource Optimization and Sustainable Development, https://www.taylorfrancis.com/chapters/edit/10.1201/9781003596721-51/research-theoretical-contributions-literature-related-tools-big-data-analytics-ravichand-kapil-bansal-lohitha-anandhi-lovi-raj-gupta-patel-chaitali-mohanbhai-narendra-kumar?context=ubx&refId=00e3f2ad-b5fc-4530-89ae-1ac3269e9566

[4] 2. E. Mythily, S. S. Ramya, K. Sangeeta, B Swathi, Manish Kumar, Purnendu_Bikash, **Narendra Kumar**: Think Big with Big Data: Finding Appropriate Big Data Strategies for Corporate Cultures, Recent Trends In Engineering and Science for Resource Optimization and Sustainable Development, https://www.taylorfrancis.com/chapters/edit/10.1201/9781003596721-46/think-big-big-data-finding-appropriate-big-data-strategies-corporate-cultures-mythily-ramya-sangeeta-swathi-manish-kumar-purnendu-bikash-narendra-kumar?context=ubx&refId=0535aba0-a7b3-4325-b543-79aa313a2168

[5] Chen, J., Song, L., & Zhou, S. (2019). Explainable AI: From black-box to interpretable machine learning models. Journal of Artificial Intelligence Research, 66(1), 1–24. https://doi.org/10.1613/jair.1.11588

[6] Guidotti, R., Monreale, A., Ruggieri, S., & Turini, F. (2018). A survey of methods for explaining black box models. ACM Computing Surveys (CSUR), 51(5), 1–42. https://doi.org/10.1145/3236009

[7] Ribeiro, M. T., & Guestrin, C. (2018). "Why should I trust you?" Explaining the predictions of black-box models. Communications of the ACM, 61(3), 56–66. https://doi.org/10.1145/3158665