International Journal of Interdisciplinary Innovative Research & Development (IJIIRD)
ISSN: 2456-236X
Vol. **11** Special Issue **02** | **2026**

A Survey of Methods, Necessity, and the Interpretability Accuracy Trade-off for Explainable AI

Sachin D. Chavhan[1], Dr. Salim Y. Amdani[2]
[1]*PhD Scholer, Department of Computer Science and Engineering, Babasaheb Naik College of Engineering Pusad, India*
[2]*Associate Professor, Department of Computer Science and Engineering, Babasaheb Naik College of Engineering, Pusad,India*

DOI: 10.5281/zenodo.19251246

ABSTRACT

As artificial intelligence systems continue to advance and permeate high-stakes application domains, the need for transparency and interpretability has become increasingly critical. Explainable AI (XAI) aims to make model predictions understandable to humans, enabling users to evaluate, trust, and manage machine-learning systems more responsibly. This paper presents a structured survey of major XAI methods, covering both model-agnostic and model-specific approaches, and examines how each contributes to improving interpretability in complex models. The necessity of explainability is discussed through practical concerns such as trust, accountability, fairness, and debugging. Furthermore, the paper analyses the interpretability–accuracy trade-off, highlighting how simpler models provide transparency while more complex models offer superior predictive performance. By organizing current methods and discussing key challenges, this survey provides a comprehensive understanding of the role of explainability in modern AI and outlines its implications for safe and effective deployment.

Keywords:- ***Explainable AI (XAI), Interpretability, Transparency, Model-Agnostic Methods, Model-Specific Methods, Interpretability-Accuracy Trade-off, Responsible AI.***

1. INTRODUCTION

Artificial intelligence (AI) and machine learning (ML) systems have achieved substantial progress in recent years, demonstrating exceptional performance in domains such as healthcare, finance, transportation, and security. This rapid adoption is largely driven by complex models particularly deep learning systems that offer high accuracy but operate as opaque “black boxes” whose internal reasoning cannot be easily understood by humans. As these models increasingly influence critical decisions, the lack of transparency raises concerns regarding trust, accountability, fairness, and safety. Users, regulators, and domain experts require explanations to verify whether an AI system’s conclusions are logical, unbiased, and consistent with real-world constraints [1], [2].

Explainable Artificial Intelligence (XAI) has emerged as a solution to this transparency barrier. XAI encompasses a wide range of techniques designed to make model predictions interpretable to humans, enabling users to understand why a certain output was produced. These explanations support essential tasks such as validating model behaviour, debugging erroneous predictions, and ensuring compliance with ethical and legal standards. As noted in recent research, interpretability is a core requirement for responsible AI deployment, especially in high-stakes settings where decisions directly impact human welfare [3], [4].

Despite significant progress, the landscape of XAI methods remains fragmented, with techniques differing in their objectives, applicability, and levels of interpretability. Existing approaches may be model-agnostic, offering explanations applicable to any classifier or regressor, or model-specific, leveraging internal model structure to produce more faithful explanations. Meanwhile, intrinsically interpretable models such as linear models or decision trees continue to offer transparency at the cost of reduced predictive accuracy. This tension highlights a longstanding challenge in machine learning: the interpretability–accuracy trade-off, where models that are easier to understand often underperform compared with more complex but opaque models [5].

The objective of this paper is to provide a structured survey of widely used XAI techniques while examining the necessity of explainability and evaluating the interpretability accuracy trade-off. By synthesizing foundational concepts and prominent methods, this survey aims to offer a clear understanding of how explainability contributes to trustworthy and reliable AI systems. The paper also discusses emerging challenges and the role of XAI in future AI system development.

2. NECESSITY OF EXPLAINABLE AI (XAI)

Explainable AI is essential as AI systems increasingly influence critical decisions in healthcare, finance, governance, and public safety. Opaque black-box models make it difficult for users to trust or verify predictions, especially when errors or biases may have serious consequences. XAI provides transparency that supports trust, accountability, fairness, and reliability in automated decision-making [2].

Explainability also enables bias detection and correction, helping practitioners identify discriminatory patterns in data or model behaviour [3]. It is equally important for safety-critical applications, where understanding AI reasoning prevents harmful or unexpected system behaviours. Emerging regulations, such as the EU GDPR and AI Act, further require AI systems to offer meaningful explanations for automated decisions [6].

Finally, XAI aids developers by providing insights into model behaviour, supporting debugging, validation, and continuous improvement. Thus, explainability is foundational for deploying AI systems that are trustworthy, legally compliant, and aligned with human reasoning.

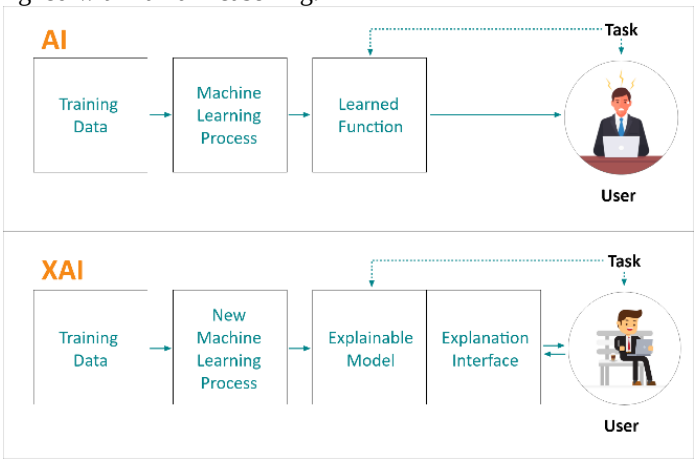

Fig.1 Explainable AI

3. XAI METHODS — A STRUCTURED SURVEY

Explainable AI (XAI) methods can be broadly categorized according to their model dependence, interpretability strategy, and scope of explanation. This section presents a structured survey of major techniques, covering model-agnostic, model-specific, post-hoc, and inherently interpretable approaches.

3.1 Model-Agnostic Explanation Methods

3.1.1 Perturbation-Based Explanation

Perturbation-based methods examine how small changes in input features influence model outputs. LIME (Local Interpretable Model-Agnostic Explanations) is one of the most prominent examples, approximating complex models with simple, interpretable surrogate models to explain individual predictions [3]. Similarly, SHAP assigns each feature a contribution value based on Shapley values from cooperative game theory and has become a standard tool due to its strong theoretical guarantees and consistency [7].

3.1.2 Example-Based Explanation

Example-based techniques explain decisions by referencing influential or similar training instances. Influence functions help estimate how individual training points affect model predictions, enabling identification of harmful or mislabelled data [8]. Counterfactual explanations provide the minimal change required in an input to obtain a different outcome, making them valuable for domains like finance and healthcare where users need actionable feedback [6].

3.2 Model-Specific Explanation Methods

Model-specific methods exploit structure unique to certain model families.

3.2.1 Trees and Ensemble Models

Decision trees are inherently interpretable, but ensemble models such as Random Forests or Gradient Boosting Machines often require explanation tools. These include feature importance scores based on impurity reduction and decision path explanations, which show how predictions follow specific tree branches.

3.2.2 Neural Network Visualization

Deep neural networks demand specialized techniques. Saliency maps highlight influential pixels based on input gradients [9]. Grad-CAM produces class-specific activation heatmaps to localize important image regions for

CNN predictions [10]. Layer-wise Relevance Propagation (LRP) distributes relevance scores backward through the network, enabling detailed attribution at the neuron level [11].

3.3 Post-Hoc Interpretability Approaches

Post-hoc approaches generate explanations after training without altering the model architecture.

3.3.1 Surrogate Models

Surrogate models approximate a complex model with a simpler interpretable one, such as a decision tree or rule list. This provides global interpretability while preserving the predictive behaviour of the original model. However, surrogate explanations may not fully reflect the underlying logic.

3.3.2 Feature Attribution Techniques

Attribution-based approaches quantify how each feature contributes to a prediction. Methods such as Integrated Gradients, Occlusion Sensitivity, and SHAP provide both local and global interpretability [7].

3.3.3 Rule Extraction

Rule extraction methods translate black-box model behaviour into human-readable IF–THEN rules. These techniques have long been used in neural network interpretability [12] and remain useful for regulatory compliance, auditing, and verification.

3.4 Inherently Interpretable Models

Unlike post-hoc approaches, inherently interpretable models incorporate transparency directly into their design.

3.4.1 Linear and Generalized Linear Models

Linear models provide coefficient-based interpretability but are limited when feature interactions are complex.

3.4.2 Decision Trees and Rule-Based Models

Tree-based models and rule lists offer clear decision pathways and remain popular in healthcare, finance, and law for their traceability.

3.4.3 Generalized Additive Models (GAMs) and Explainable Boosting Machines (EBMs)

GAMs represent decisions as sums of interpretable feature functions. Explainable Boosting Machines (EBMs) extend this concept using boosted trees to achieve near state-of-the-art accuracy while retaining full transparency [13].

3.4.4 Prototype-Based Networks

Models such as ProtoPNet compare inputs with learned prototypes, aligning predictions with human reasoning by providing example-based explanations.

4. INTERPRETABILITY–ACCURACY TRADE-OFF

A central challenge in the deployment of machine learning systems is the perceived trade-off between interpretability and predictive accuracy. In many domains, highly expressive models such as deep neural networks, ensemble-based boosting, and transformer architectures achieve state-of-the-art performance but remain opaque and difficult to explain. Conversely, simple and transparent models like linear regression or decision trees provide clear reasoning pathways but may underperform in complex, high-dimensional problem spaces [5]. As a result, practitioners often face a dilemma: whether to prioritize model interpretability or model accuracy.

The trade-off becomes most evident in safety-critical and regulated applications. For example, in medical diagnosis or criminal risk assessment, stakeholders require not only high prediction accuracy but also clarity regarding the factors influencing outcomes. A highly accurate model that cannot justify its decision may be unsuitable regardless of performance metrics if it prevents clinicians, legal authorities, or affected individuals from understanding assessment criteria [2]. In such cases, a slightly less accurate but interpretable model may be preferred because transparency improves trust, accountability, and legal compliance.

However, the trade-off is not always inevitable. Recent research has demonstrated the potential for transparent yet high-performing models. Explainable Boosting Machines (EBMs), for instance, retain the interpretable structure of Generalized Additive Models while achieving prediction accuracy competitive with black-box systems across multiple domains [13]. Similarly, post-hoc interpretability tools such as SHAP and Grad-CAM enable practitioners to gain insights into complex models without altering their architecture [7][10]. These approaches suggest that interpretability and accuracy can be complementary rather than mutually exclusive when applied strategically.

Ultimately, the interpretability accuracy trade-off reflects a broader decision-making challenge: selecting the right model for the right application. In low-risk, high-performance-driven environments, opaque models may

be acceptable. In contrast, in high-risk domains where fairness, transparency, and user understanding are essential, interpretability becomes a primary design requirement. Progress in XAI research continues to narrow this gap, enabling the development of systems that are both accurate and accountable.

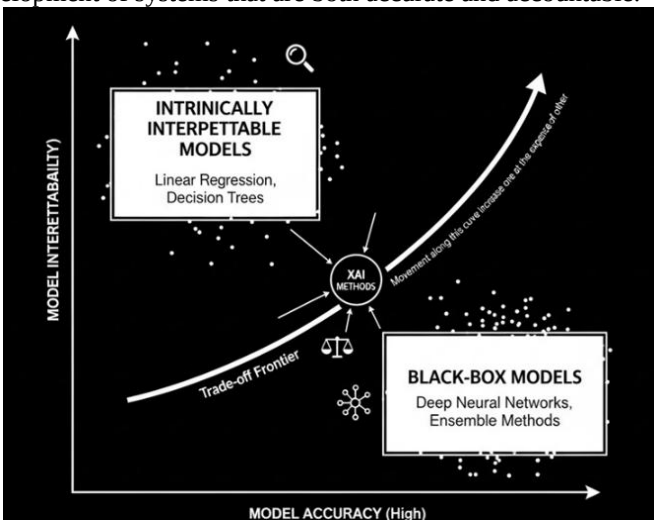

Fig.2 Interpretability Accuracy Trade-off diagram

5. CHALLENGES

Despite rapid advancements in XAI, several unresolved challenges continue to limit its seamless integration into real-world AI systems. One of the biggest issues is the lack of a standard measurement of interpretability. Current interpretability metrics such as fidelity, completeness, and compactness are neither universally defined nor consistently applied across models and domains, making comparisons difficult and slowing scientific progress [2]. Moreover, many widely adopted explanation techniques (e.g., LIME, SHAP) operate as post-hoc approximations, leading to concerns about whether explanations truly reflect the model's underlying reasoning or merely summarise it [3].

Another major challenge involves the interpretability–accuracy trade-off. While inherently interpretable models promote transparency, they often underperform on complex datasets compared to black-box deep learning architectures. Therefore, organisations must choose between optimal performance and explainability, raising critical questions about fairness, usability, and risk in high-stakes environments such as healthcare and finance [14].

The usability gap also remains significant. Most explanation tools are designed for technical experts and are not intuitive for end-users such as doctors, loan officers, or policymakers. To truly democratise AI, explanations must be adaptive to users' cognitive needs, enabling non-technical stakeholders to reason confidently with AI recommendations [1]. Additionally, XAI must evolve beyond trust and transparency toward Human–AI cognitive alignment, where explanations bridge differences between human reasoning and machine learning decision processes enabling collaborative, not merely observable, AI.

From an engineering standpoint, XAI faces major scalability difficulties: producing explanations for large-scale deep learning models is computationally expensive, and real-time explainability is still impractical in many scenarios. There are also rising concerns regarding privacy leakage through explanations, as saliency maps or feature-influence estimates may unintentionally reveal sensitive training data [15].

6. CONCLUSION

Explainable AI (XAI) has emerged as a critical research area aimed at addressing the limitations of opaque machine learning systems and enabling responsible AI adoption across high-stakes domains. This survey reviewed the necessity of XAI, highlighted key motivation factors, and provided a structured overview of major interpretability techniques including model-agnostic, model specific, intrinsically interpretable, and deep learning–based explanation approaches. The discussion also examined the interpretability accuracy trade-off, emphasising the ongoing tension between performance and transparency in real-world deployments.

Although XAI has achieved significant progress, current methods still face challenges related to standardisation, evaluation complexity, user-centric design, scalability, and risks such as privacy leakage. The future of XAI depends on a transition from post-hoc explanations toward holistic transparency integrating interpretability with fairness, safety, robustness, causality, and human-AI cognitive alignment. To unlock the full potential of AI in

sensitive industries such as healthcare, law, finance, and autonomous systems, explainability must evolve from a supplemental feature to a fundamental design principle.

Ultimately, advancing XAI is not just a technical objective but a societal necessity. As AI systems continue to influence human decisions, progress in explainability will play a crucial role in ensuring accountability, trust, and collaboration between humans and intelligent machines paving the way for safe, ethical, and sustainable AI adoption.

7. REFERENCES

[1] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," Artificial Intelligence, vol. 267, pp. 1–38, 2019.

[2] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv:1702.08608, 2017.

[3] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," in Proc. 22nd ACM SIGKDD, 2016, pp. 1135–1144.

[4] C. Molnar, Interpretable Machine Learning, 2nd ed. 2022.

[5] Z. C. Lipton, "The mythos of model interpretability," Communications of the ACM, vol. 61, no. 10, pp. 36–43, 2018.

[6] Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a Right to Explanation Does Not Exist in the GDPR. International Data Privacy Law.

[7] Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. NIPS.

[8] Koh, P. W., & Liang, P. (2017). Understanding Black-box Predictions via Influence Functions. ICML.

[9] Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps.

[10] Selvaraju, R. R., et al. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. ICCV.

[11] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W.(2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLOS ONE.

[12] Andrews, R., Diederich, J., & Tickle, A. B. (1995). Survey and Critique of Techniques for Extracting Rules from Trained Artificial Neural Networks. Knowledge-Based Systems.

[13] Caruana, R., Lou, Y., et al. (2015). Intelligible Models for Healthcare. KDD.