iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** II **Month of publication:** February 2025

DOI: https://doi.org/10.22214/ijraset.2025.66968

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue II Feb 2025- Available at www.ijraset.com

Explainability in AI: Interpretable Models for Data Science

Ali Mohammed Omar Ali[1], Abbas Abubaker Mohammed Ahessin[2], Moussa Mohamed Naji Haqi[3]
[1,2]*College of Technical Sciences – Sebha*
[3]*Higher institute of Technical Engineering - Sebha*

Abstract: ***As artificial intelligence (AI) continues to drive advancements across various domains, the need for explainability in AI models has become increasingly critical. Many state-of-the-art machine learning models, particularly deep learning architectures, operate as "black boxes," making their decision-making processes difficult to interpret. Explainable AI (XAI) aims to enhance model transparency, ensuring that AI-driven decisions are understandable, trustworthy, and aligned with ethical and regulatory standards. This paper explores different approaches to AI interpretability, including intrinsically interpretable models such as decision trees and logistic regression, as well as post-hoc methods like SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations). Additionally, we discuss the challenges of explainability, including the trade-off between accuracy and interpretability, scalability issues, and domain-specific requirements. The paper also highlights real-world applications of XAI in healthcare, finance, and autonomous systems. Finally, we examine future research directions, emphasizing hybrid models, causal explainability, and human-AI collaboration. By fostering more interpretable AI systems, we can enhance trust, fairness, and accountability in data science applications.***
Keywords: Explainable AI (XAI), Interpretability, Machine Learning, Black-Box Models, Model Transparency, SHAP, LIME, Ethical AI, Trustworthy AI, Post-hoc Explainability, Bias Mitigation, Regulatory Compliance, Human-AI Interaction.

I. INTRODUCTION

Explainability in Artificial Intelligence (AI) refers to the ability of an AI system to provide clear and understandable reasons for its decisions and predictions. As AI models, particularly those based on machine learning and deep learning, become increasingly complex, their decision-making processes often become opaque, leading to what is commonly referred to as "black-box" models. This lack of transparency can be problematic, especially in critical applications such as healthcare, finance, and criminal justice, where understanding the rationale behind a decision is crucial for trust, accountability, and ethical considerations [1].

Interpretable models in data science are designed to address this issue by providing insights into how inputs are transformed into outputs. These models are structured in a way that their internal workings can be easily understood by humans. For instance, linear regression models, decision trees, and rule-based systems are inherently interpretable because they follow a clear and logical process that can be easily visualized and explained.

The importance of explainability in AI cannot be overstated. It enables stakeholders to validate the model's decisions, identify potential biases, and ensure compliance with regulatory requirements. Moreover, explainability fosters user trust and facilitates the adoption of AI technologies across various domains [2].

Artificial Intelligence (AI), particularly machine learning (ML) models, has achieved remarkable success in various fields, including healthcare, finance, and autonomous systems. However, many of these models, especially deep learning-based architectures, function as "black boxes," making it difficult to understand their decision-making processes. Explainability in AI (XAI) seeks to bridge this gap by developing methods that make models more interpretable and transparent

This lack of transparency raises concerns about trust, fairness, and accountability, especially in high-stakes applications such as medical diagnostics, fraud detection, and autonomous driving. Discovering patterns and structures in large troves of data in an automated manner is a core component of data science, and currently drives applications in diverse areas such as computational biology, law and finance. However, such a highly positive impact is coupled with significant challenges: how do we understand the decisions suggested by these systems in order that we can trust them? Indeed, when one focuses on data-driven methods—machine learning and pattern recognition models in particular—the inner workings of the model can be hard to understand. In the very least, explainability can facilitate the understanding of various aspects of a model, leading to insights that can be utilized by various stakeholders, such as (Figure 1):

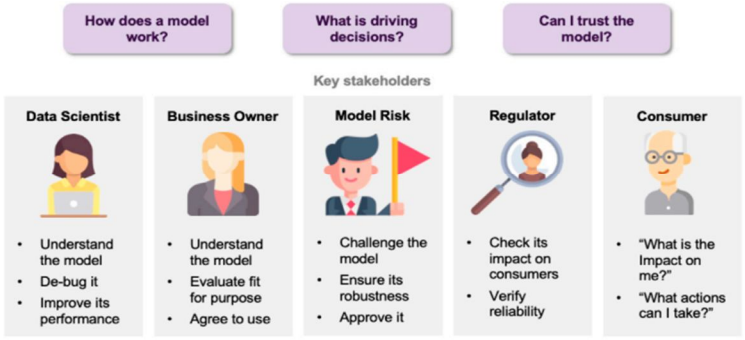


Figure 1 : Concerns faced by various stakeholders.

II. INTRINSICALLY INTERPRETABLE MODELS

A range of atomistic ML models has been introduced in recent years [2]. The focus has mainly been on the regression of atom-resolved properties, or global properties as dependent on individual atomic environments. The construction of structural descriptors is often guided by physical ideas, encoding information about environments and symmetries, but this is noting indispensable practice, as complex neural networks have also been used to capture materials structures from raw data inputs. The former naturally lend themselves to interpretable models . The development of physically motivated interatomic potentials from machine learning has been comprehensively covered in other review articles.[3]

We finish dealing with intrinsically interpretable models by noting that it is also important not to fetishize simpler models in the name of interpretability. Particularly important in this regard is the scenario of model mismatch, where the model form fails to capture the true form of a relationship; i.e., according to our previous definition, it provides low correctness. For example, if a linear model is used to capture a nonlinear relationship, the model will increasingly attribute importance to irrelevant features in an attempt to minimize the difference between the model predictions and the training data and will ultimately produce meaningless explanations. In machine learning literature, a common solution to preserve predictive power and allow high intrinsic interpretability is using generalized linear models with specific linkage functions or generalized additive models (GAMs) [4]. For example, GAMs have been used to model and interpret the driving factors of chemical adsorption of subsurface alloys, 17 modeling a nonlinear process with a high degree of interpretability. Linear models are not always as interpretable as they claim to be. For example, if features are heterogeneous and have very different ranges and values, the coefficients of a linear model will probably tell us more about the sizes of various parameters than they will tell us about some underlying physical explanation that is understandable to a domain expert [5].

III. MODEL EXPLANATION METHODS

Though some ML methods offer intrinsically interpretable results, many more complex models such as deep neural networks (DNNs) are not as easily understood. Even when models are inherently interpretable, extrinsic interpretation methods can provide additional insights impossible by examining the model alone. What-If Explanations There are a range of "what-if" analyses that work by examining how the value of the model output changes when one or more of the input values are modified. Partial dependence plots (PDPs)examine how changing a given feature affects the output, ignoring the effects of all other features. For example, we could look at the effect of the mean atomic mass of a material on the dielectric response, marginalizing all other factors using a model such as that presented in ref 5. One drawback is that confounding relationships are missed and can mask effects [6].

Explainable AI (XAI) has emerged as a crucial field of research aimed at making AI models more understandable, interpretable, and trustworthy. The goal of XAI is to provide meaningful explanations for AI-driven decisions, enabling users—whether domain experts, regulators, or end-users—to comprehend why a model made a particular prediction. Explainability not only enhances trust in AI systems but also facilitates bias detection, regulatory compliance, and ethical decision-making.

There are two primary ways to achieve explainability in AI:

1) Intrinsic Interpretability: Some models, such as decision trees and logistic regression, are inherently interpretable due to their simple structure. These models allow users to directly understand how inputs contribute to outputs.

2) Post-hoc Explainability: For complex models like deep neural networks and ensemble methods, interpretability techniques such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) help explain model predictions after training.

Despite advancements in XAI, several challenges remain, including the trade-off between accuracy and interpretability, scalability issues in large AI models, and the difficulty of creating explanations that are both precise and human-understandable. Moreover, different domains require different levels of explainability—what is interpretable to a data scientist may not be comprehensible to a medical practitioner or a financial analyst. A number of classifications of XAI techniques for deep learning models have been proposed [6]. Drawing on this work, XAI techniques can be classified using two dimensions:
(i) whether the technique is model-specific or model-agnostic and (ii) whether the technique is designed to provide an explanation that is global in scope to the model or one that is local in scope to a prediction (Table 1).

Table 1: Classification of XAI Techniques

	Model-specific	Model-agnostic
Global	Enforce interpretability constraints into the structure and learning mechanisms of deep learning models	Develop interpretable global surrogate models based on input-output associations predicted by a black-box model Apply diagnostic techniques to understand the importance of specific features in a black-box model's predictions
Local	Use attention mechanisms to show how the model selectively focuses on features in high-dimensional input for an instance	Develop interpretable surrogate models with local fidelity in the vicinity of an instance

This paper explores the importance of explainability in AI, different interpretability techniques, current challenges, and real-world applications in fields such as healthcare, finance, and autonomous systems. Finally, we discuss future research directions, including hybrid models, causal explainability, and the integration of explainability tools in AI development pipelines [8].
This explainability requirement lead a new area of AI research, know as Explainable AI (XAI). (Figure 2) shows how XAI can add new dimensions to AI by answering the "wh" questions that were missing in traditional AI. The XAI, therefore, has drawn a great interest from critical applications, such as health care, defence, law and order, etc., where explaining how an answer was obtained (i.e., answers to "wh" questions) is as important as obtaining the answer. In both academia and industry, XAI research, therefore, has become a priority [9]. Although a number of work have been already proposed, more and more work is required to realize the full potential of XAI.

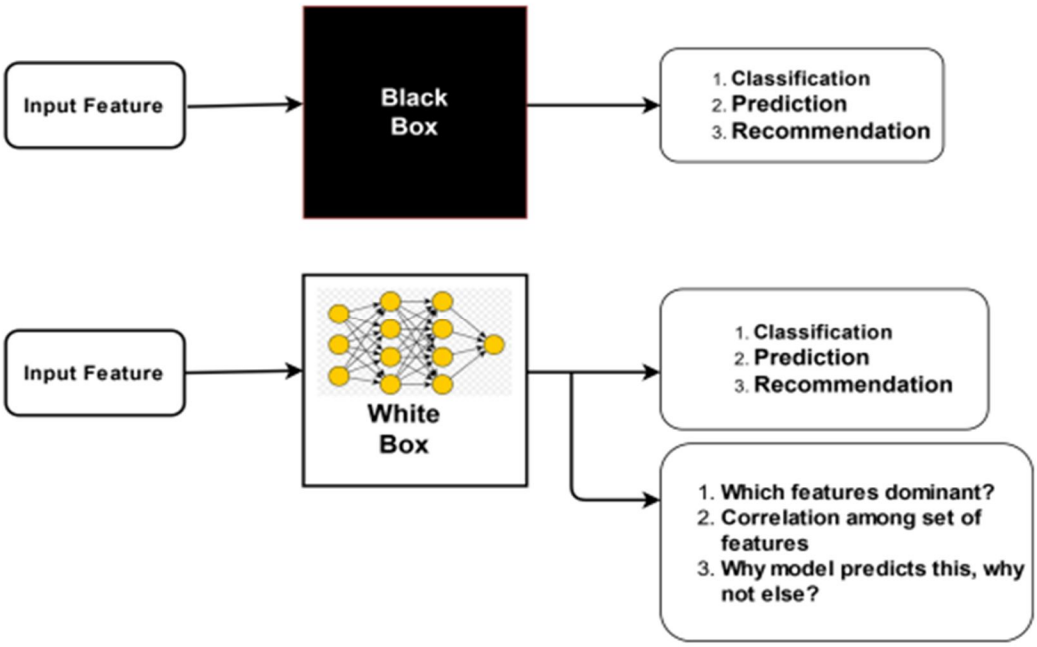


Figure 2 : AI vs XAI

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue II Feb 2025- Available at www.ijraset.com

IV. LITERATURE SYNTHESIS

A thorough analysis of the subject of explanation would have to cover literatures spanning the history of Western philosophy: Disciplines including philosophy and psychology of science, cognitive psychology, psycholinguistics, and expert systems.

Considering xAI as a metahuman system - a unique kind of sociotechnical systems [10 we examined the articles with a sociotechnical lens in mind [11]. A sociotechnical approach takes a holistic view where relations among people, technology, tasks and organization are sustainable. From previous IS research we know that "poorly designed sociotechnical systems with inadequate concern with mutual relationships were shown to fail and produce unintended or unwanted outcomes" [11, p. 8]. Taking this as a departure point we examine the article pool if concerns like a) a holistic view of social and technical aspect are considered in the xAI literature; b) consideration or participation of relevant stakeholders in xAI design, development and use processes.

have developed 'PIRL', a Programmatically Interpretable Reinforcement Learning framework, as an alternative to DRL. In DRL, the policies are represented by neural networks, making them very hard (if not impossible) to interpret. The policies in PIRL, on the other hand, while still mimicking the ones from the DRL model, are represented using a high-level, human-readable programming language. Here, the problem stays the same as in traditional RL (i.e., finding a policy that maximises the long-term reward), but in addition, they restrict the vast amount of target policies with the help of a (policy) sketch. To find these policies, they employ a framework which was inspired by imitation learning, called Neurally Directed Program Search (NDPS). This framework first uses DRL to compute a policy which is used as a neural 'oracle' to direct the policy search for a policy that is as close as possible to the neural oracle [12].

A conceptual model of the XAI explaining process is presented in (Figure 3). This diagram highlights four major classes of measures. Initial instruction in how to use an AI system will enable the user to form an initial mental model of the task and the AI system. Subsequent experience, which can include system-generated explanations, would enable to participant to refine their mental model, which should lead to better performance and appropriate trust and reliance [13].

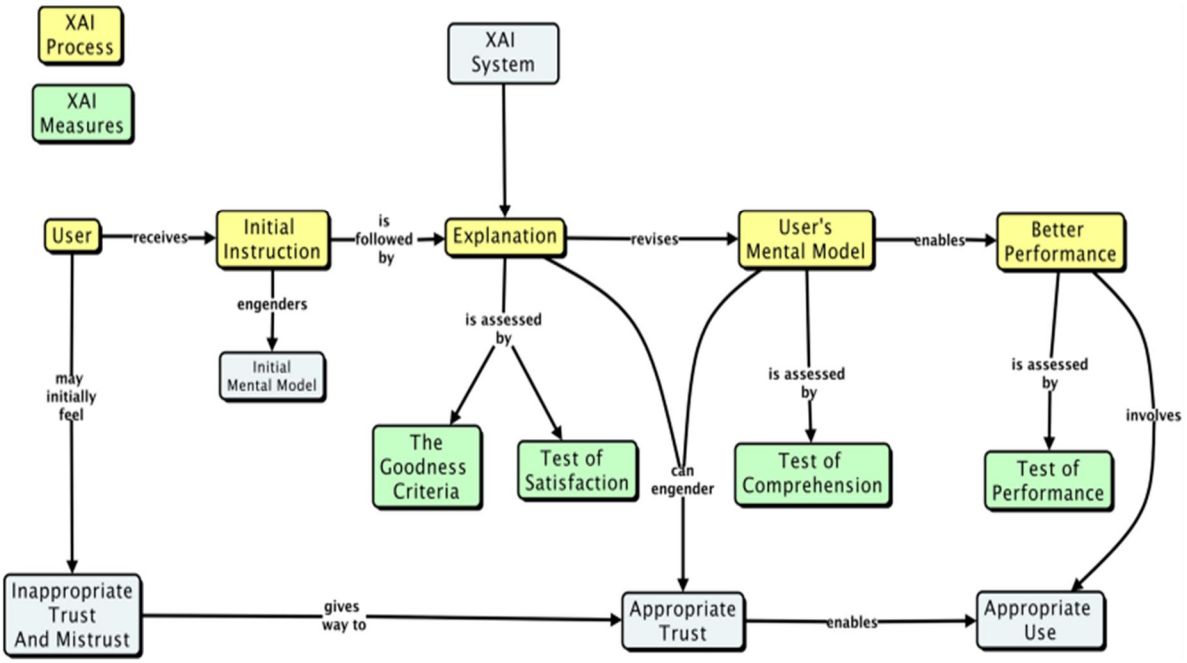


Figure 3: A conceptual model of the process of explaining, in the XAI context.

V. IMPORTANCE OF EXPLAINABILITY IN AI

1) Trust & Transparency: Users and stakeholders must understand how AI models arrive at decisions to trust their outputs.
2) Regulatory Compliance: Industries like finance and healthcare require AI explainability to comply with regulations such as GDPR and HIPAA.
3) Bias & Fairness Detection: Understanding model behavior can help identify and mitigate biases in training data.
4) Debugging & Performance Improvement: Explainability aids in diagnosing errors and refining models [14].

VI. CATEGORIES OF INTERPRETABLE MODELS

AI models can be categorized based on their interpretability:

1) Intrinsic Interpretability: Some models are inherently interpretable due to their simple structure, including:
   - Decision Trees: Provide rule-based explanations for decisions.
   - Linear & Logistic Regression: Coefficients provide insight into feature importance.
   - k-Nearest Neighbors (k-NN): Classification is based on similarity to nearby data points.
2) Post-hoc Explainability: Some models require external techniques to explain their behavior:
   - Feature Importance Analysis: Identifies which features contribute most to predictions.
   - Local Explanations: Examines individual predictions rather than the entire model.

VII. METHODS FOR EXPLAINABILITY IN AI

Several approaches help interpret black-box AI models:

1) SHAP (Shapley Additive Explanations): Based on cooperative game theory, it assigns importance scores to features.
2) LIME (Local Interpretable Model-agnostic Explanations): Generates simpler models to approximate a complex model locally.
3) Attention Mechanisms in Deep Learning: Highlights relevant parts of input data that influence the model's decision.
4) Counterfactual Explanations: Shows what changes in input would lead to different outcomes.
5) Gradient-Based Methods (e.g., Grad-CAM): Visualizes important features in neural networks[15].

VIII. CHALLENGES IN AI EXPLAINABILITY

1) Trade-off Between Accuracy and Interpretability: More interpretable models often have lower predictive performance.
2) Scalability Issues: Some explainability techniques are computationally expensive for large models.
3) Domain-Specific Interpretability Needs: Explanations must be tailored to specific applications, such as healthcare or finance.
4) User Understanding: Explanations must be intuitive for non-technical users [16].

IX. APPLICATIONS OF EXPLAINABLE AI

Explainable AI (XAI) plays a crucial role across various domains where AI-driven decisions impact human lives, financial systems, security, and regulatory compliance. Below are key application areas where explainability is essential:

1) Healthcare: AI-driven diagnostics must provide justifications for predictions to assist doctors in decision-making.
2) Finance: Loan approvals and fraud detection systems must explain their reasoning to ensure fairness.
3) Autonomous Systems: Self-driving cars require explainable AI to enhance safety and reliability.
4) Legal & Ethical AI: Explainability helps organizations ensure ethical AI deployment.

X. FUTURE DIRECTIONS

1) Hybrid Models: Combining interpretable models with black-box techniques for better transparency.
2) Human-AI Collaboration: Developing interactive explainability tools for better user engagement.
3) Causal Explainability: Moving beyond correlation-based explanations to causal reasoning [17,18,19].

Explainability in AI is essential for building trust, ensuring fairness, and meeting regulatory requirements across various industries. From healthcare and finance to cybersecurity and self-driving cars, XAI enhances transparency, making AI-driven decisions more understandable and accountable. As AI adoption grows, research in explainability techniques like SHAP, LIME, and counterfactual explanations will become increasingly important. Future developments in human-centered AI, hybrid models, and ethical guidelines will further improve XAI applications, ensuring AI systems remain fair, reliable, and transparent [20].

XI. CONCLUSION

As artificial intelligence (AI) becomes increasingly integrated into critical sectors such as healthcare, finance, cybersecurity, and autonomous systems, the need for explainability in AI has grown significantly. Many state-of-the-art AI models, particularly deep learning architectures, operate as black-box systems, making their decision-making processes opaque. This lack of transparency raises concerns related to trust, accountability, fairness, and regulatory compliance.

Explainable AI (XAI) addresses these challenges by providing methods to interpret, understand, and justify AI-driven decisions. This paper explored various approaches to AI interpretability, including intrinsic interpretability (models such as decision trees and logistic regression) and post-hoc explainability techniques (such as SHAP, LIME, Grad-CAM, and counterfactual explanations).

While these methods improve transparency, they also present challenges such as computational complexity, the trade-off between accuracy and interpretability, and domain-specific requirements.

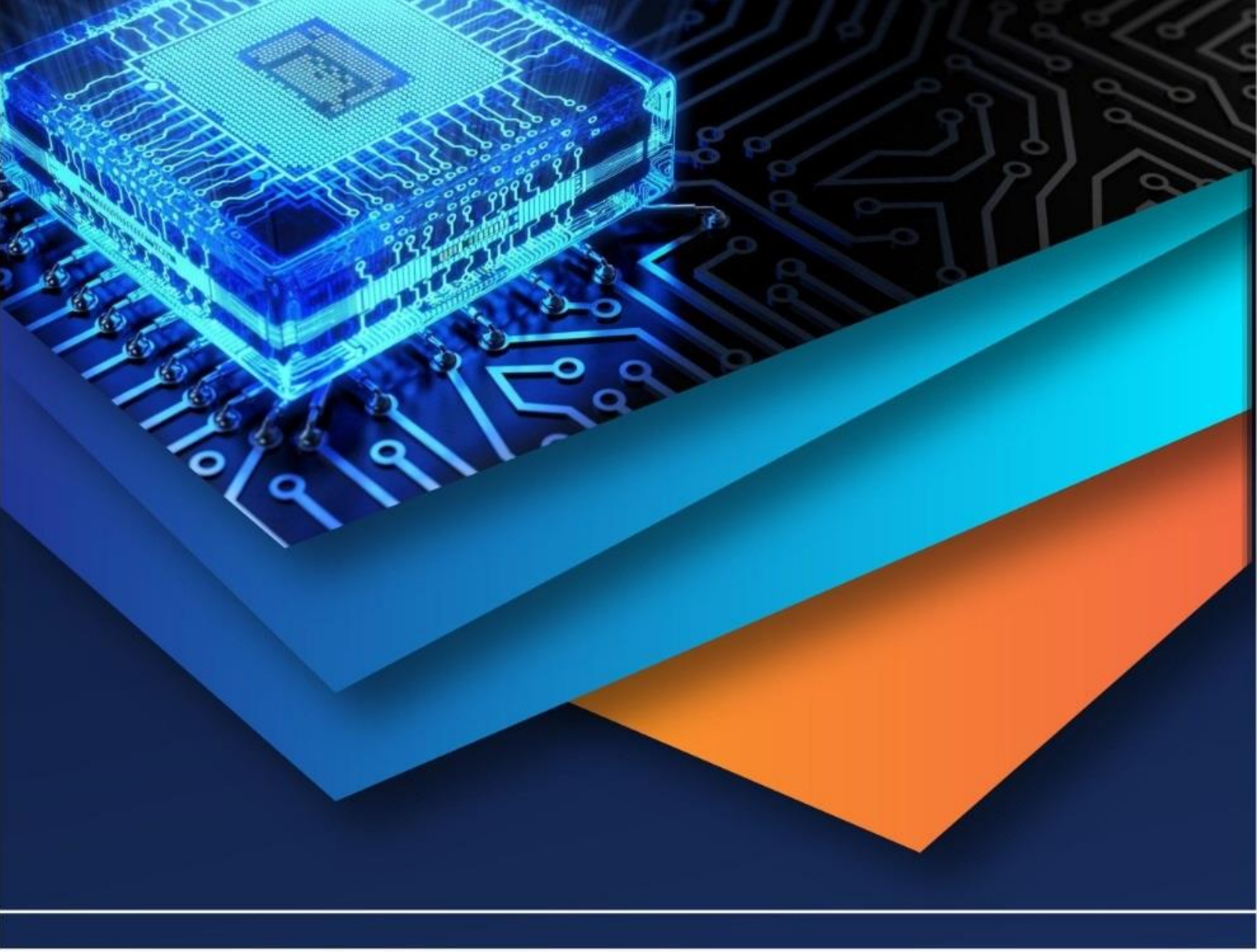
The applications of XAI span multiple industries, including healthcare (AI-driven diagnostics and personalized medicine), finance (fraud detection and credit scoring), autonomous systems (self-driving cars and robotics), cybersecurity (threat detection and intrusion prevention), and policy-making (legal AI and ethical decision-making). In each of these fields, explainability is not just an added feature—it is a necessity to ensure that AI models are reliable, unbiased, and aligned with ethical and legal standards.

Despite significant progress, several open challenges remain. AI models must balance high predictive performance with transparency, and explanations should be both technically accurate and easily understandable for diverse stakeholders. Future research directions in XAI include developing hybrid models that integrate interpretable AI components with deep learning, advancing causal explainability methods, and enhancing human-AI collaboration to build trustworthy AI systems. Additionally, as regulatory bodies introduce new AI governance frameworks, organizations must incorporate explainability as a core component of AI deployment.

In conclusion, explainability is fundamental to the responsible and ethical use of AI. By continuing to refine XAI methods and integrating them into AI systems, researchers and practitioners can build AI technologies that are transparent, accountable, and beneficial to society. The future of AI lies not only in making accurate predictions but also in ensuring that these predictions are understood, trusted, and ethically sound.

REFERENCES

[1] Athey, S., Imbens, G. W. 2015 Machine-learning methods https://arxiv.org/abs/1504.01132v1 (see also ref. 7).
[2] Caruana, R., Kangarloo, H., Dionisio, J. D, Sinha, U., Johnson, D. 1999. Case-based explanation of non-casebased learning methods. In Proceedings of the American Medical Informatics Association (AMIA) Symposium: 212-215.
[3] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the 21st Annual SIGKDD International Conference on Knowledge Discovery and Data Mining, 1721-1730.
[4] Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., Blei, D. M. 2009. Reading tea leaves: how humans interpret topic models. In Proceedings of the 22nd International Conference on Neural Information Processing Systems (NIPS), 288-296.
[5] Amershi, S., Chickering, M., Drucker, S.M., Lee, B., Simard, P., & Suh, J. (2015). Modeltracker: Redesigning performance analysis tools for machine learning. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (pp. 337–346).
[6] New York: Association for Computing Machinery. Biran, O., & Cotton, C. (2017). Explanation and Justification in Machine Learning: A Survey. IJCAI-17 Workshop on Explainable Artificial Intelligence (XAI). Brézillon, P., & Pomerol, J.-C. (1997).
[7] Joint cognitive systems, cooperative systems and decision support systems: A cooperation in context. In Proceedings of the European Conference on Cognitive Science, Manchester (pp. 129–139). Chi, M.T., Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994).
[8] Eliciting self-explanations improves understanding. Cognitive Science, 18(3), 439–477. Clancey, W.J. (1986a). From GUIDON to NEOMYCIN and HERACLES in twenty short lessons. AI Magazine, 7(3), 40. Doshi-Velez, F., & Kim, B. (2017).
[9] A Roadmap for a Rigorous Science of Interpretability. ArXiv Preprint ArXiv:1702.08608. Retrieved from https://arxiv.org/abs/1702.08608 Goodman, B., & Flaxman, S. (2016).
[10] European Union regulations on algorithmic decision-making and a “right to explanation.” Presented at the ICML Workshop on Human Interpretability in Machine Learning, New York, NY. Johnson, H., & Johnson, P. (1993).
[11] Explanation Facilities and Interactive Systems. In Proceedings of the 1st International Conference on Intelligent User Interfaces (pp. 159–166).
[12] New York: Association for Computing Machinery. Kass, R., & Finin, T. (1988). The Need for User Models in Generating Expert System Explanation. International Journal of Expert Systems, 1(4), 345–375. Krull, D. S., & Anderson, C.
[13] Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of Inmates Running the Asylum.
[14] In Proceedings of the International joint Conference on Artificial Intelligence (IJCAI-17) Workshop on Explainable Artificial Intelligence (XAI). Moore, J. D., & Swartout, W. R. (1991).
[15] A reactive approach to explanation: taking the user’s feedback into account. In C. Paris, W.R. Swartyout, & W.C.
[16] Mann (Eds.), Natural language generation in artificial intelligence and computational linguistics (pp. 3–48). New York: Springer. Mueller, S.T. & Klein, G. (March-April 2011). Improving users’ mental models of intelligent software tools. IEEE: Intelligent Systems, 26(2), 77—83.
[17] Arrieta, A.B., Rodríguez, N.D., Ser, J.D., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F., 2020.
[18] Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 58, 82–115. URL: https://doi.org/10. 1016/j.inffus.2019.12.012, doi:10.1016/j.inffus.2019.12.012.
[19] Austin, J., Urmson, J., Sbisà, M., 1975. How to Do Things with Words. William James lectures, Clarendon Press. URL: https: //books.google.it/books?id=XnRkQSTUpmgC.
[20] Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., Schneider, N., 2013. Abstract meaning representation for sembanking, in: Dipper, S., Liakata, M., Pareja-Lora, A. (Eds.), Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, LAWID@ACL 2013, August 8-9, 2013, Sofia, Bulgaria, The Association for Computer Linguistics. pp. 178–186. URL: https://aclanthology. org/W13-2322/.
[21] Zou X, Yang J, Zhang H, Li F, Li L, Wang J, et al. (2023). Segment Everything Everywhere All at Once. In: Advances in Neural Information Processing Systems (A Oh, T Naumann, A Globerson, K Saenko, M Hardt, S Levine, eds.), volume 36, 19769–19782. Curran Associates, Inc.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY