A Review of XAI Models and Applications: Recent Developments and Future Trends

Prof. Ashwini G. Mote[1]
[1]*Assistant Professor, Department of Computer Science & Engineering, VVPIET, Soregaon, Solapur, India*

Abstract— **Explainable Artificial Intelligence (XAI) has become increasingly indispensable as AI systems permeate diverse sectors, necessitating transparency and interpretability for user trust and regulatory compliance. This systematic review paper comprehensively examines recent developments and future trends in XAI models and applications. Artificial Intelligence (AI) uses systems and machines to simulate human intelligence and solve common real-world problems. Machine learning and deep learning are Artificial intelligence technologies that use algorithms to predict outcomes more accurately without relying on human intervention. However, the opaque black box model and cumulative model complexity can be used to achieve. Explainable Artificial Intelligence (XAI) is a term that refers to Artificial Intelligence (AI) that can provide explanations for their decision or predictions to human users. XAI aims to increase the transparency, trustworthiness and accountability of AI system, especially when they are used for high-stakes application such as healthcare, finance or security. This paper offers systematic literature review of XAI approaches with different application**

Key words—**Machine Learning, Deep Learning, Explanation, Explainable AI, Healthcare**

I.INTRODUCTION

As AI technologies are increasingly deployed across diverse and high-stakes domains such as healthcare, finance, legal decision-making, and autonomous systems, the traditional "black-box" nature of many advanced ML models poses significant challenges. These challenges include a lack of trust from users, difficulties in regulatory compliance, and potential ethical issues arising from biased or opaque decision-making processes.

XAI aims to bridge the gap between complex ML models and human understanding by developing methods and tools that make these models' decisions comprehensible. This is crucial for ensuring that AI systems are not only accurate but also trustworthy, fair, and accountable. Recent years have seen substantial progress in XAI, with advancements in model-specific explainability methods, post-hoc interpretability techniques, and their application in various sectors.

Humans are naturally capable of thinking and performing certain tasks on their own due to brain. Human brain has the natural ability to perform some analytical tasks such as object recognition much faster than any computer. This has inspired researchers and scientist to build machines that can do similar tasks designed by human brain. Humans and animals have the ability to naturally learn, remember, make decisions and perform many complex tasks and this cognitive capability is called Natural Intelligence. While humans and animals used the fuzzy (approximate reasoning) logic to learn and make decisions, computer machines are developed to do the same tasks using crisp (binary) logic. This process of developing intelligence using computers or similar machines is called Artificial Intelligence (AI). The ultimate goal of AI is to build a machine that can think and act like a human and automate complex tasks which can be performed efficiently. There are several branches of AI that makes a system comparable to Human intelligence and builds AI powered systems to digitize common mundane tasks and eliminate repetitive tasks.

Explainable artificial intelligence (XAI) has been proposed as a solution that can help to move towards more transparent AI and thus avoid limiting the adoption of AI in critical domains [1], [2]. Generally speaking, according to [3], XAI focuses on developing explainable techniques that empower end-users in comprehending, trusting, and efficiently managing the new age of AI systems.

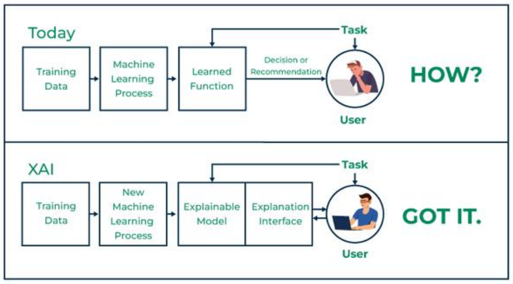

Figure-Explainable AI Concept

A.What is Explainable AI?

Explainable artificial intelligence (XAI) refers to a collection of procedures and techniques that enable machine learning algorithms to produce output and results that are understandable and reliable for human users. Explainable AI is a key component of the fairness, accountability, and transparency (FAT) machine learning paradigm and is frequently discussed in connection with deep learning. Organizations looking to establish trust when deploying AI can benefit from XAI. XAI can assist them in comprehending the behavior of an AI model and identifying possible problems like AI.

B. Why Explainable AI is needed?

Nowadays, we are surrounded by black-box AI systems utilized to make decisions for us, as in autonomous vehicles, social networks, and medical systems. Most of these decisions are taken without knowing the reasons behind these decisions. The need for explainable AI arises from the fact that traditional machine learning models are often difficult to understand and interpret. These models are typically black boxes that make predictions based on input data but do not provide any insight into the reasoning behind their predictions. This lack of transparency and interpretability can be a major limitation of traditional machine learning models and can lead to a range of problems and challenges.[3]

One major challenge of traditional machine learning models is that they can be difficult to trust and verify. Because these models are opaque and inscrutable, it can be difficult for humans to understand how they work and how they make predictions. This lack of trust and understanding can make it difficult for people to use and rely on these models and can limit their adoption and deployment.

II. XAI Models

1. Model-Specific Methods

- Interpretable Models: Algorithms inherently designed to be interpretable, such as decision trees, linear models, and rule-based systems, have seen continued improvements in terms of accuracy and scalability. Advances include enhanced pruning techniques for decision trees and more efficient algorithms for generating rule sets.
- Self-Explaining Models: These models, such as attention mechanisms in neural networks, provide internal transparency by highlighting which parts of the input data are most influential in the decision-making process. Recent work has focused on making these models more robust and their explanations more meaningful.

2. Post-Hoc Explainability:

- Feature Importance: Techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) have been refined to provide more accurate and computationally efficient explanations. Enhancements include better sampling methods for LIME and improved estimation techniques for SHAP values.
- Visual Explanations: Methods like Grad-CAM (Gradient-weighted Class Activation Mapping) and saliency maps have been further developed to provide clearer and more interpretable visual explanations for image-based models.

III. APPLICATIONS OF XAI

1. Healthcare:

- Diagnostics and Prognostics: XAI models are being used to interpret complex medical data, aiding in diagnosing diseases and predicting patient outcomes. For example, interpretable models can help clinicians understand the basis for diagnostic predictions, leading to better-informed decisions.[4]

- Personalized Medicine: By explaining the factors influencing treatment recommendations, XAI helps in tailoring treatments to individual patients, enhancing the effectiveness of personalized medicine.

2. Finance:

- Credit Scoring: XAI models help in explaining credit scores and loan approval decisions, which is crucial for regulatory compliance and customer trust. They ensure that decisions are fair and transparent, reducing the risk of biased outcomes.
- Fraud Detection: In fraud detection, XAI techniques help in understanding and interpreting anomalous patterns in transaction data, enabling quicker and more accurate identification of fraudulent activities.

3. Legal and Ethical Decision-Making:

- Fairness and Bias Detection: XAI methods are employed to detect and mitigate biases in AI systems, ensuring fairer outcomes across different demographic groups. This is particularly important in legal settings where decisions can have significant social implications.
- Transparent Decision Processes: In applications like parole decisions and sentencing, XAI provides transparency, ensuring that decisions are based on understandable and justifiable factors.

4. Autonomous Systems:

- Safety and Trust: For autonomous vehicles and drones, XAI helps in understanding and validating the decision-making processes, which is crucial for safety and gaining public trust.
- Human-Robot Interaction: In collaborative robotics, XAI enables robots to explain their actions to human partners, improving coordination and efficiency.

IV. BENEFITS OF EXPLAINABLE AI

The value of explainable AI lies in its ability to provide transparent and interpretable machine-learning models that can be understood and trusted by humans. This value can be realized in different domains and applications and can provide a range of benefits and advantages. Some of the key values of explainable AI include:

- Improved decision-making:– Explainable AI can provide valuable insights and information that can be used to support and improve decision-making. For example, explainable AI can provide insights into the factors that are most relevant and influential in the model's predictions, and can help to identify and prioritize the actions and strategies that are most likely to achieve the desired outcome.
- Increased trust and acceptance:– Explainable AI can help to build trust and acceptance of machine learning models, and can overcome the challenges and limitations of traditional machine learning models, which are often opaque and inscrutable. This increased trust and acceptance can help to accelerate the adoption and deployment of machine learning models and can provide valuable insights and benefits in different domains and applications.
- Reduced risks and liabilities:– Explainable AI can help to reduce the risks and liabilities of machine learning models, and can provide a framework for addressing the regulatory and ethical considerations of this technology. This reduced risk and liability can help to mitigate the potential impacts and consequences of machine learning, and can provide valuable insights and benefits in different domains and applications.

V. HOW DOES EXPLAINABLE AI WORK?

The architecture of explainable AI depends on the specific approaches and methods that are used to provide transparency and interpretability in machine learning models. However, in general, explainable AI architecture can be thought of as a combination of three key components:

- Machine learning model:– The machine learning model is the core component of explainable AI, and represents the underlying algorithms and techniques that are used to make predictions and inferences from data. This component can be based on a wide range of machine learning techniques, such as supervised, unsupervised, or reinforcement learning, and can be used in a range of applications, such as medical imaging, natural language processing, and computer vision.

- Explanation algorithm:- The explanation algorithm is the component of explainable AI that is used to provide insights and information about the factors that are most relevant and influential in the model's predictions. This component can be based on different explainable AI approaches, such as feature importance, attribution, and and visualization, and can provide valuable insights into the workings of the machine learning model.
- Interface:- The interface is the component of explainable AI that is used to present the insights and information generated by the explanation algorithm to humans. This component can be based on a wide range of technologies and platforms, such as web applications, mobile apps, and visualizations, and can provide a user-friendly and intuitive way to access and interact with the insights and information generated by the explainable AI system.

VI. EXPLAINABLE AI PRINCIPLES

Explainable AI (XAI) principles are a set of guidelines and recommendations that can be used to develop and deploy transparent and interpretable machine learning models. These principles can help to ensure that XAI is used in a responsible and ethical manner, and can provide valuable insights and benefits in different domains and applications. Some of the key XAI principles include:

1. Transparency:- XAI should be transparent and should provide insights and information about the factors that are most relevant and influential in the model's predictions. This transparency can help to build trust and acceptance of XAI and can provide valuable insights and benefits in different domains and applications.
2. Interpretability:– XAI should be interpretable and should provide a clear and intuitive way to understand and use the insights and information generated by XAI.
3. Accountability:– XAI should be accountable and should provide a framework for addressing the regulatory and ethical considerations of machine learning. This accountability can help to ensure that XAI is used in a responsible and accountable manner, and can provide valuable insights and benefits in different domains and applications.

VII. FUTURE TRENDS

1. Integration of XAI with AI Development:

- XAI will become an integral part of the AI development lifecycle, from model training to deployment, ensuring transparency and accountability at every stage.

2. Standardization and Benchmarking:

- Development of standardized metrics and benchmarks for evaluating the quality and effectiveness of explanations will be essential for advancing the field.

3. Human-Centered Design:

- Greater focus on designing XAI methods that cater to the needs and cognitive capabilities of end-users, ensuring that explanations are not just technically sound but also comprehensible and actionable.

4. Interdisciplinary Research:

- Collaboration between AI researchers, domain experts, and social scientists will be crucial for addressing the complex challenges of explainability, particularly in high-stakes domains like healthcare and law.

5. Regulatory and Ethical Frameworks:

- The evolution of regulatory and ethical frameworks will guide the development and deployment of XAI, ensuring that AI systems are transparent, fair, and accountable.

VII. XAI TECHNIQUES

XAI techniques are diverse and continuously evolving, aimed at enhancing the transparency and interpretability of machine learning models. Explainable AI (XAI) encompasses a variety of techniques designed to make the outputs of machine learning models understandable to humans. These techniques can be broadly categorized into model-specific methods, post-hoc explainability methods, and approaches involving causal inference and counterfactuals.

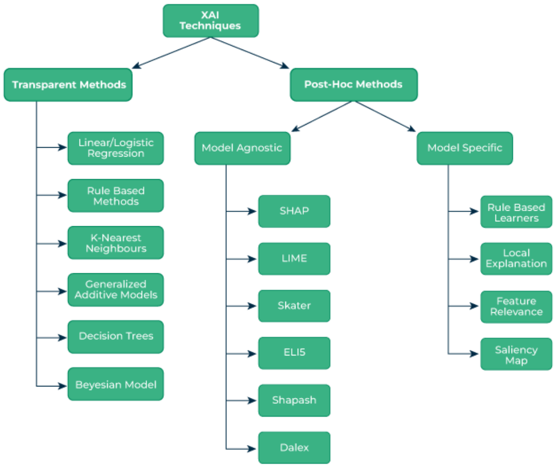

Figure-XAI Techniques

VIII. CONCLUSION

Explainable AI is a rapidly evolving field with significant advancements in model-specific methods, post-hoc explainability, and applications across various domains. The future of XAI lies in its integration into the AI development process, standardization of evaluation metrics, and a human-centered approach to design. As AI systems become more pervasive, the importance of explainability will continue to grow, ensuring that these systems are transparent, trustworthy, and aligned with human values.

REFERENCE

[1] Adadi A., Berrada M.Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)IEEE Access, 6 (2018), pp. 52138 52160, 10.1109/ACCESS.2018.2870052

[2] Barredo Arrieta A., Díaz-Rodríguez N., Del Ser J., Bennetot A., Tabik S., Barbado A., Garc ia GilLopez S., Molina D., Benjamins R., Chat ila R., Herrera F.Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI Inf. Fusion, 58 (2020), pp. 82-115, 10.1016/j.inffus.2019.12.012

[3] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, 1135-1144.

[4] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82-115.

[5] 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 (NIPS'17), 4768-4777.

[6] K.Chen, *LinearNetworksandSystems*, Belmont, CA:Wadsworth,1993, pp. 123-135. Markus A.F., Kors J.A., Rijnbeek P.R. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies.

[7] Burkart N., Huber M.F.A survey on the explainability of supervised machine learningJ. Artificial Intelligence Res., 70 (2021), pp. 245-317, 10.1613/jair.1.12228

[8] Guidotti R., Monreale A., Ruggieri S., Turini F., Giannotti F., Pedreschi D. A survey of methods for explaining black box models ACM Comput. Surv., 51 (5) (2018), 10.1145/3236009

[9] Payrovnaziri S.N., Chen Z., Rengifo- Moreno P., Miller T., Bian J., Chen J.H., Liu X., He Z. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review J. Am. Med. Inform. Assoc., 27 (7) (2020), pp.1173-1185, 10.1093/ jamia/ ocaa053.

[10] Samek W., Müller K.-R. Towards explainable artificial intelligence Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer International Publishing, Cham (2019), pp. 5-22, 10.1007/978-3-030-28954-6_1

[11] O. Biran, C. Cotton, Explanation and justification in machine learning: A survey, in:IJCAI-17 Workshop on Explainable AI, Vol. 8,XAI, (1) 2017, pp. 8–13.

[12] D. Gunning, Broad Agency Announcement Explainable Artificial Intelligence (XAI), Technical report, 2016.