Second International Conference on Smart Computing and Communication for Sustainable Convergence (ISCCSC 2025)

A Survey on Explainable Artificial Intelligence(XAI) Techniques

Anusha I M[1], Srushti Umarani[2], and Kasilingam N[3]

Department of Artifial Intelligence and Machine Learning, Alva's institute of Engineering And Technology, Mijar,Moodbidri – 574225, Mangalore, Karnataka, India

Email:anushamulimani2004@gmail.com, srushtipu14@gmail.com, kasilingamn@gmail.com

Abstact

Due to the growing influence of AI on decisions in various industries like healthcare, finance, and autonomous systems as well as criminal justice, transparency and trustworthiness in AI are becoming increasingly important considerations. While state-of-the-art AI systems (deep learning models for example) are often referred to as 'black box' algorithms because of how they function, Explainable-AI or XAI aims to address this challenge through providing people with insight and understanding into how AI has arrived at its conclusions so as to enable users to make decisions based on their confidence in the decisions made by AI. This paper will provide a detailed review of all explanations generated from XAI techniques; organized according to either model-intrinsic or post hoc approaches to establishing XAI. Examples of commonly used methods/approaches include but are not limited to: LIME; SHAP; Grad-CAM; Attention Mechanisms; Expert Knowledge-Based Rule Models; etc., with each example illustrating both the pros/cons of these various methods along with application areas where they are most useful. Further, this paper will also give an overview of the different evaluation metrics used to measure XAI effectiveness, as well as current issues and areas needing continued research effort in the field of XAI.

Keywords. Accelerometer, gyroscope, GPS tracking, machine learning, smart walking stick, fall detection, emergency alarm system, ultrasonic sensor, IoT, assistive technology.

1. INTRODUCTION

Machine learning and deep learning have made great strides in the effectiveness of artificial intelligence (AI). Still, the intricate workings of AI create a situation in which it can be very challenging for users to understand how AI reaches its conclusions. This challenge can cause users significant risk in many use cases where human accountability, fairness, and safety are of utmost importance. Because many regulators have implemented regulations like the

General Data Protection Regulation (GDPR) that grant individuals the right to know how automated decisions are made, these regulations highlight the need for transparency in AI's decision-making processes. XAI has been developed to provide users with explanations that allow for the reconciliation of the model's proficiency and a user's ability to interpret the model's reasoning using an approach that is understandable, faithful, and actionable.

2. LITERATURE REVIEW

2.1.Interpretable models (White-Box Models)
Interpretable or white-box models (such as linear regression, decision trees and rule-based systems) are transparent by design, and their inherent simplicity makes them easy for users to see how input features affect their predictions. Despite their transparency, interpretable models typically cannot achieve the same level of predictive accuracy and effectiveness as more complex deep learning architectures can.
2.2. Models using Post Hoc Explainability Techniques
Post hoc explainability techniques are applied after the training of the model has been completed, and LIME generates explanations for local decision boundaries using interpretable surrogate models and SHAP uses game theory to estimate feature importance. Visualization-based explainability techniques, such as Grad-CAM, identify regions of interest (the most salient regions) and indicate the areas that were most responsible for a prediction made by a convolutional neural network model.
2.3. Model-Specific Techniques of Explainability
Some explainability methods are based on a given class of models. For example, attention mechanisms within neural networks can be used to identify the parts of an input to which a model directs its focus during a prediction process. Decision trees, on the other hand, have their own definition of importance and can be analyzed by the user based on both feature importance metrics and decision path analysis.

3. CLASSIFICATION OF XAI TECHNIQUES

3.1. Intrinsically Understandable Models
The first category of methodologies are models that have natural interpretable structures such as Generalized Additive Models (GAM), Decision Trees (DT) and Rule-Based Classifiers. These models were built with a focus on transparency, which means they are naturally interpretable and therefore provide insight into their predictions. They do however sacrifice some accuracy for interpretability when it comes to more complicated datasets.
3.2. Post-Hoc Explainability
This category represents methods to explain the predictions from black box (or opaque) machine learning models after they have been created. This category has a large degree of flexibility; it can be used with nearly any type of model and as a result, it has a high likelihood for producing estimation errors and providing less believable results compared to Intrinsically Understandable Models.

4. EVALUATION METRICS

The evaluation of methods in explainable artificial intelligence (XAI) is difficult because of the subjective nature of how people perceive things. Currently, common criteria by which we evaluate answers from machine learning methods to explain why decisions were made

are fidelity, interpretability, completeness, stability, and user trust. Human-centered evaluations are becoming increasingly prevalent in validating whether machine learning methods are actually providing useful explanations, and these include both user studies and expert assessments.

5. APPLICATIONS

XAI has been implemented in a variety of fields, including Healthcare — Clinical Decision Support (CDS), Finance — Credit Scoring and Fraud Detection, Autonomous Vehicles — Vehicle Safety Assurance, Cybersecurity — Cyber Threat Analysis.

6. CHALLENGES

XAI has made great strides but is still limited by issues like finding a balance between accuracy and interpretability, developing a set of standard methods to evaluate models with XAI, large models pose challenges for scalability, and sometimes we can overpromote the results of the implementations from XAI as they can potentially mislead users. Many researchers are currently looking for ways to address all of these limitations.

7. FUTURE DIRECTIONS

Future research in the areas of XAI will include future work on creating hybrid explainability methods, developing standardized benchmark methods, format and define domain-specific explanation frameworks, and incorporating a human-centered design approach. A major area of focus will be on creating a way for XAI to be integrated with fairness, robustness, and privacy in developing responsible AI solutions.

6. CONCLUSION

Explainable Artificial Intelligence is a critical component of ensuring transparency and building trust and accountability in modern AI systems. The survey examined major XAI techniques and their classification systems, applications, and limitations. As AI continues to evolve, XAI will be a foundation of responsible and ethical AI deployment.

7. REFERENCES

[1] A. Adadi and M. Berrada, “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence,” IEEE Access, 2018.

[2] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why Should I Trust You?” Explaining the Predictions of Any Classifier, KDD, 2016.

[3] S. M. Lundberg and S. I. Lee, “A Unified Approach to Interpreting Model Predictions,” NeurIPS, 2017.

[4] R. R. Selvaraju et al., “Grad-CAM: Visual Explanations from Deep Networks,” ICCV, 2017.

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

Biographies

Anusha I M is a 4th-year student studying AIML at Alva's Institute of Engineering and Technology (AIET), and throughout her time at AIET, has developed a strong interest inML, Data science and Data Analytics.

Srushti Umarani is also a fourth-year student studying AIML at AIET, and has a particular interest in LLM and Gen Models, as well as working with innovative technologies.

Kasilingam N is an Assistant Professor within the AIML Department at AIET. He graduated with a B.Tech degree followed by a Master of Engineering degree from Sona College of Engineering and Technology, located in Salem, Tamil Nadu.