International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com | A Peer-Reviewed, Refereed, a Bimonthly Journal|

|| Volume 8, Issue 2, March – April 2025 ||

DOI: 10.15680/IJCTECE.2025.0802016

Explainable Artificial Intelligence: A Comprehensive Review of Methods, Applications, Challenges, and Future Research Directions

Sneha Valusa

AI/ML Consultant, USA

snehavalusa@outlook.com

Sankethu Babu Guntaka

AI Security Consultant, USA

sankethuguntaka@outlook.com

Venkata Phanindra Lingam

AI Architect, USA

venakataphanindra@gmail.com

ABSTRACT: Explainable Artificial Intelligence is an important area of research which addresses the lack of transparency of current artificial intelligence systems, particularly those utilizing deep learning and other sophisticated machine learning algorithms. This research paper presents an exhaustive review of Explainable Artificial Intelligence by discussing its past history, basic concepts, taxonomy, state-of-the-art explanation techniques, applications and research challenges. First, this article discusses the history of the development of AI from naturally interpretable machine learning algorithms to complex black-box deep learning architectures to highlight the increasing requirement of explainability. This article then goes ahead and categorizes the XAI techniques into intrinsic/post-hoc, local/global explanations and model-specific/model-agnostic techniques. In this research, we provide a critical review of popular explanation techniques including Local Interpretable Model-agnostic Explanations, Gradient-weighted Class Activation Mapping, Saliency Maps, Integrated Gradients, Attention Visualization, Counterfactual Explanations, Rule-based Explanation, Decision Tree and Feature Importance techniques in terms of their basic principles, advantages, disadvantages and applicability to various AI models.

KEYWORDS: Explainable Artificial Intelligence, Explainable AI, LIME, SHAP, Grad-CAM, Human-Centered AI.

I. INTRODUCTION

Artificial Intelligence has transformed many areas of science, industry and social life by enabling computers to see, think, learn, forecast and make decisions in a human-like way. Machine learning and deep learning are rapidly evolving and improving intelligent systems for medical diagnosis, financial forecasting, cybersecurity, autonomous mobility, natural language processing, computer vision, and software engineering. But even with these huge advances, many current AI models are black-box systems, and the process of internal decision making is incomprehensible to humans. This opacity has inhibited the use of AI in critical applications where decisions directly affect human lives, financial assets, public safety and legal accountability.

The growing complexity of deep neural networks, transformer architectures and large-scale foundation models has generated problems of explainability, fairness, accountability, robustness and ethical AI deployment. Users, subject experts and regulatory authorities typically want accurate projections and comprehensive explanations of how and why an AI system takes a choice. Doctors need interpretable explanations before they accept AI-assisted medical diagnoses, financial institutions need transparent credit-scoring models to satisfy regulatory requirements, and cybersecurity experts need to grasp AI-generated threat assessments before they take protective steps. Explainability of AI decisions is a fundamental need for trustworthy human-centric intelligent systems.

| ISSN: 2320-0081 | www.ijctece.com | A Peer-Reviewed, Refereed, a Bimonthly Journal|

|| Volume 8, Issue 2, March – April 2025 ||

DOI: 10.15680/IJCTECE.2025.0802016

Explainable Artificial Intelligence (XAI) is an interdisciplinary research field that enhances the transparency and interpretability of AI models, without compromising their prediction accuracy. XAI offers a variety of techniques to help humans understand, validate, and trust AI decisions. Explanations may be generated by means of intrinsic interpretable models, post-hoc explanations, local or global interpretation methods, and model-specific or model-agnostic approaches, depending on the application and model complexity. Popular methods for interpretation of sophisticated machine learning and deep learning models across domains include LIME, SHapley Additive Explanations, Gradient-weighted Class Activation Mapping (Grad-CAM), Integrated Gradients, Saliency Maps, Counterfactual Explanations and Attention Visualisation.

The main contributions of this survey may be summarised as follows:
1. It covers the development and basic principles of Explainable Artificial Intelligence from classical interpretable models to recent deep learning systems.
2. It offers a systematic taxonomy of XAI based on intrinsic and post-hoc explainability, local and global explanations, and model-specific and model-agnostic techniques.
3. Critically evaluates the popular explainability approaches such as LIME, SHAP, Grad-CAM, Saliency Maps, Integrated Gradients, Counterfactual Explanations, Rule-based Explanations, Decision Trees and Feature Importance Approaches.

II. RELATED WORKS

One of the first important contributions was that of Ribeiro et al . (2016) with the Local Interpretable Model-Agnostic Explanations framework. LIME [4] proposed a model-agnostic technique to explain individual predictions by learning an interpretable surrogate model in a small neighbourhood of the input instance. This work showed that LIME is one of the most popular local explanation methodologies that can be used to interpret proper black-box models without any modification to the underlying design.

Lundberg and Lee (2017) then proposed SHapley Additive exPlanations, a game theoretic approach that provides a coherent framework for computing feature importance based on the classic Shapley values from cooperative game theory. SHAP delivers local and global explanations and fulfils desirable theoretical properties like consistency and local accuracy. SHAP is a standard way to evaluate machine learning and deep learning models, due to its strong mathematical foundation and broad use.

Gradient-weighted Class Activation Mapping was suggested by Selvaraju et al. (2017) in the context of computer vision. It gives visual heatmaps of the picture regions important for the prediction of the convolutional neural network. Grad-CAM was a major step forward in the explainability of CNN-based image classification models and is currently a widespread visualisation approach used in medical imaging, autonomous driving and item recognition.

Many scholars have also researched feature attribution methods based on gradients. Sundararajan et al. (2017) proposed Integrated Gradients, a technique that attributes the prediction of a model to the input features by integrating the gradients along a path from a baseline input to the original input. Most importantly, Integrated Gradients satisfies fundamental axioms like sensitivity and implementation invariance that justify explanations over standard saliency methods.

International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com | A Peer-Reviewed, Refereed, a Bimonthly Journal|

|| Volume 8, Issue 2, March – April 2025 ||

DOI: 10.15680/IJCTECE.2025.0802016

III. ARCHITECTURE DIAGRAM

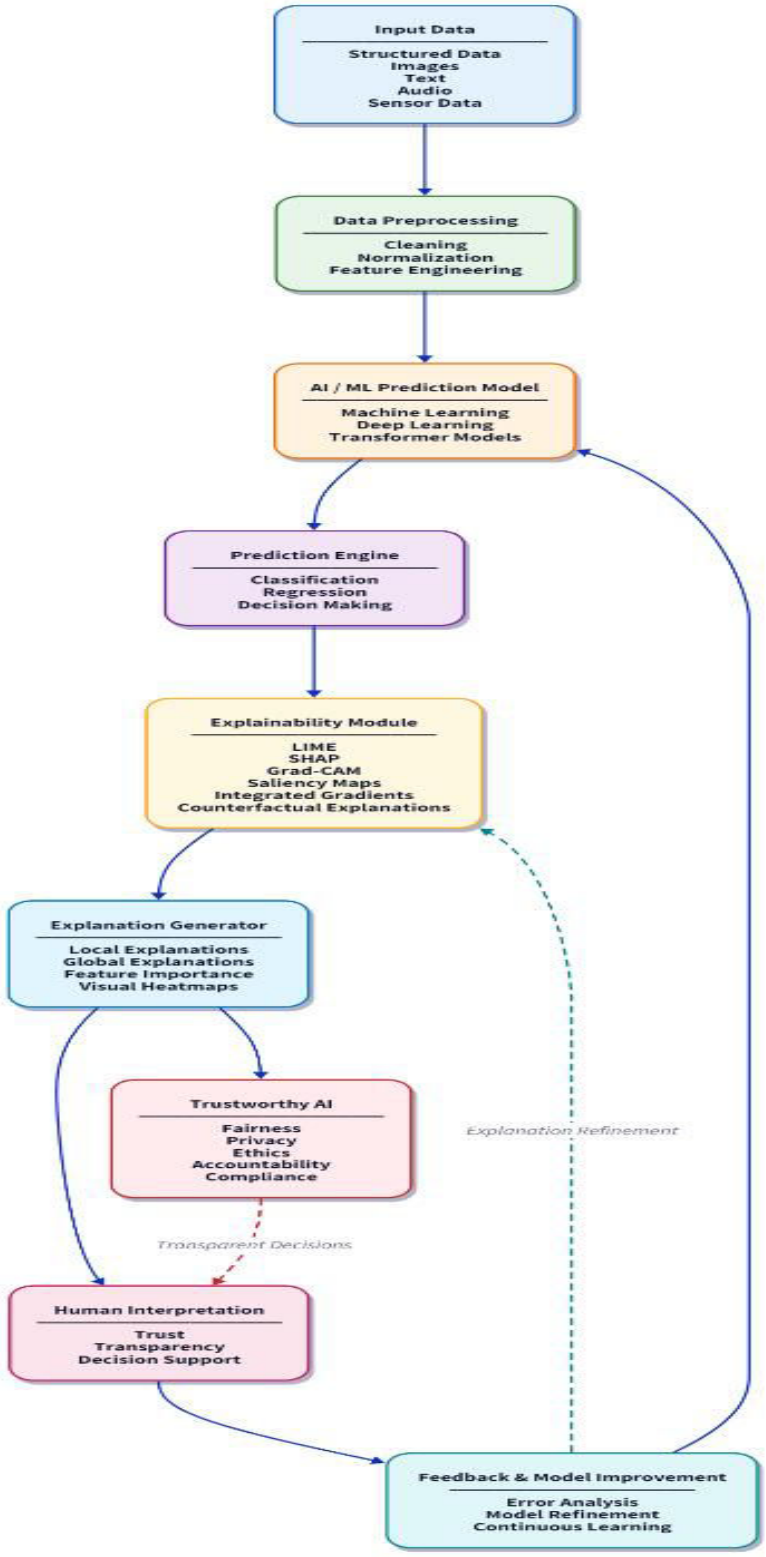


Fig 1: Architecture Diagram

| ISSN: 2320-0081 | www.ijctece.com | A Peer-Reviewed, Refereed, a Bimonthly Journal|

|| Volume 8, Issue 2, March – April 2025 ||

DOI: 10.15680/IJCTECE.2025.0802016

IV. EXPLAINABILITY TECHNIQUES

Explainable Artificial Intelligence (XAI) uses many methods to make sense of the predictions of machine learning and deep learning models. These strategies are different in terms of methodology, scope, application and level of interpretability. XAI approaches may be broadly classified into model-agnostic or model-specific, local or global, and intrinsic or post-hoc explanations. We explore the most popular explainability approaches, their concepts, benefits, limits, and application fields.

4.1 Local Interpretable Model-Agnostic Explanations

Local Interpretable Model-Agnostic Explanations by Ribeiro et al. (2016) is one of the first and most important post-hoc explainability methodologies. LIME provides an interpretation for the prediction of any black box model by approximating it locally around a given instance with a simple interpretable model such as linear regression or a decision tree. Instead of providing the whole model, LIME generates local explanations, which describe the most relevant variables for a particular prediction. LIME is a model-agnostic tool that may be used to classification, regression, image analysis and natural language processing problems. LIME is computationally fast and simple to comprehend. However, the explanations generated by LIME might vary based on the sampling approach and neighbourhood selection, resulting to instability in particular applications.

4.2 SHapley Additive exPlanations

SHAP is a unified paradigm for explainability introduced by Lundberg and Lee (2017) on the basis of cooperative game theory. It computes Shapley values to quantify the contribution of each input feature to a model prediction. SHAP has desired theoretical qualities such as local correctness, consistency and missingness, making it one of the most dependable explanation techniques accessible. SHAP may offer local explanations of individual forecasts and global explanations of overall feature relevance, unlike LIME. SHAP is a technique rooted in solid mathematics that has gained popularity in healthcare, finance, cybersecurity, and industrial analytics. However, correct calculation of the Shapley values may be computationally costly for big datasets and sophisticated models.

Table 1. Comparison of Explainability Techniques

Technique	Category	Explanation Scope	Model Type	Advantages	Limitations
LIME	Post-hoc	Local	Model-agnostic	Simple, interpretable, flexible	Explanation instability
SHAP	Post-hoc	Local & Global	Model-agnostic	Strong theoretical foundation	High computational cost
Grad-CAM	Post-hoc	Local	CNN	Visual heatmaps	Limited to CNNs
Saliency Maps	Post-hoc	Local	Deep Learning	Easy implementation	Sensitive to noise
Integrated Gradients	Post-hoc	Local	Deep Learning	Reliable feature attribution	Baseline selection required
Attention Visualization	Post-hoc	Local	Transformers	Interprets attention mechanisms	Attention may not fully explain reasoning
Counterfactual Explanations	Post-hoc	Local	Model-agnostic	Actionable explanations	Computationally complex
Rule-based Explanations	Intrinsic/Post-hoc	Global	Various Models	Human-readable rules	Difficult for complex models
Decision Trees	Intrinsic	Global	Decision Tree	Naturally interpretable	Lower accuracy for complex problems
Feature Importance	Post-hoc	Global	Most ML Models	Identifies influential features	Model-dependent interpretation

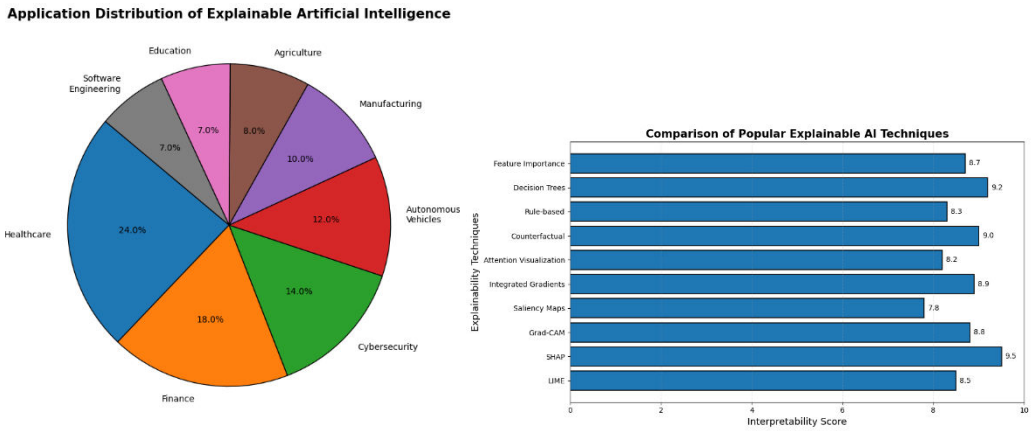
International Journal of Computer Technology and Electronics Communication (IJCTEC)

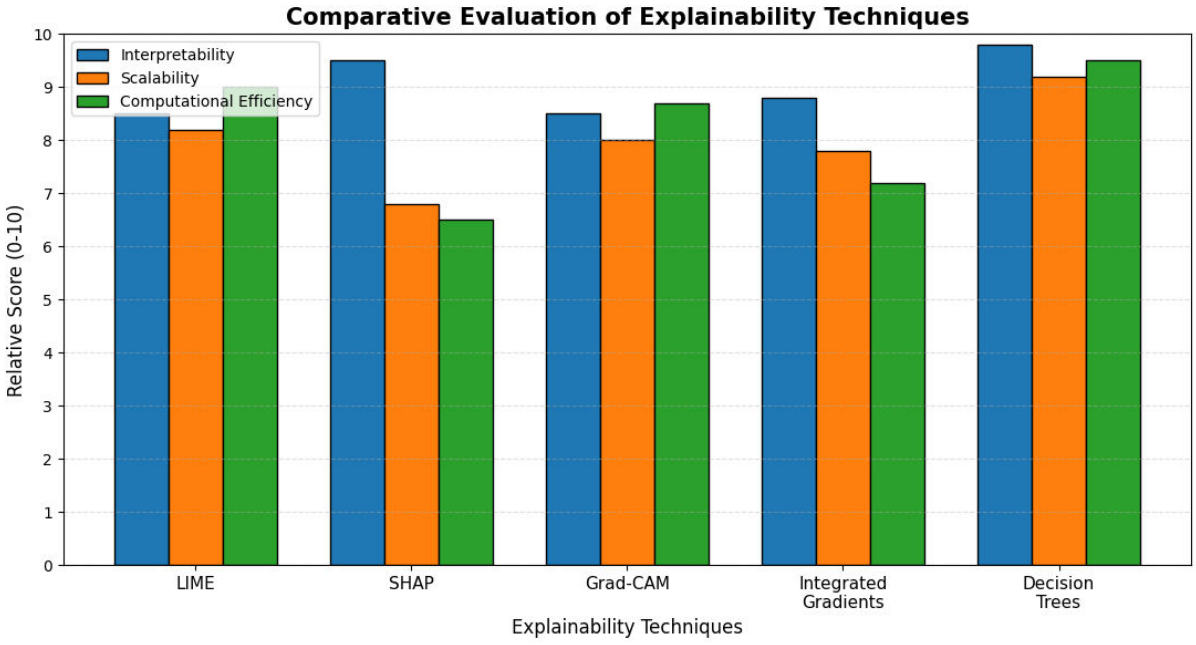
| ISSN: 2320-0081 | www.ijctece.com | A Peer-Reviewed, Refereed, a Bimonthly Journal|

|| Volume 8, Issue 2, March – April 2025 ||

DOI: 10.15680/IJCTECE.2025.0802016

V. VISUALIZATION GRAPH





VI. CHALLENGES

Despite considerable advancements achieved by Explainable Artificial Intelligence, there are still a number of technological and practical barriers that stand against its ubiquitous deployment. The first barrier refers to the problem of the tradeoff between interpretation and predictive performance, as highly interpretable models tend to exhibit worse results compared to more complex deep learning models. Besides, many explainability techniques are not scalable to large datasets and foundation models due to increased computing complexity. Some other critical issues relate to algorithmic biases and fairness, as well as privacy protection due to the potential disclosure of confidential information and encouragement of biased decision-making. In addition, another relevant research problem involves the issue of explanation resilience and robustness under adversarial conditions. Finally, it is necessary to ensure human interpretability and interpretability to diverse stakeholders, including domain experts and non-experts, which requires human interpretability. Finally, the lack of standardized criteria for measuring explainability solutions' performance prevents objective evaluation of their quality, validity and usefulness.

VII. LIMITATION

Explainable Artificial Intelligence has helped achieve great transparency in modern-day AI technologies, there still are many problems. First, some techniques do not provide exact, but approximate explanations which do not always

| ISSN: 2320-0081 | www.ijctece.com | A Peer-Reviewed, Refereed, a Bimonthly Journal|

|| Volume 8, Issue 2, March – April 2025 ||

DOI: 10.15680/IJCTECE.2025.0802016

correspond to the reasoning of complex models. Second, some explainability methods, like SHAP and Integrated Gradients, consume much computational power, preventing their use in time- and resource-constrained scenarios. Also, different explainability techniques sometimes provide different and even conflicting explanations for one and the same prediction, which can affect users' trust. Besides, the majority of existing XAI tools is aimed at working with specific types of models or specific application domains, and thus generalizing them to all AI technologies is problematic.

VIII. CONCLUSION

XAI has come out as a crucial part of AI by overcoming the problem of transparency with respect to machine learning and deep learning models. This survey has focused on discussing the history, core concepts, taxonomy, methods, applications, and challenges of XAI, which has become increasingly important in various fields including healthcare, finance, cyber security, autonomy, manufacturing, education, and software engineering. XAI makes it possible for the end-users to make sure of the decisions made by AI algorithms and build confidence in their outcomes. Although there have been significant advancements in this field, reaching a compromise among explainability, prediction accuracy, computation time, and fairness still poses a research challenge ahead.

IX. FUTURE WORKS

Potential avenues for future research in Explainable Artificial Intelligence are developing better and more scalable ways to provide explanations that will help to support even more complex AI models, such as foundation models and large language models. Some examples are the integration of causal explanations, multimodal explanations, private XAI, federated XAI, and human-in-the-loop learning to create transparency without reducing the performance of models and compromising data security. Creating standardized ways to measure and evaluate the quality of explanations is crucial for comparing the effectiveness of different techniques and ensuring responsible use of AI technologies. Finally, advancements in explainable reinforcement learning, autonomous agents, and trustworthy generation of AI can extend the applications of XAI.

REFERENCES

[1] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
[2] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144).
[3] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2921–2929).
[4] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (Vol. 30, pp. 4765–4774).
[5] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (pp. 618–626).
[6] Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning (pp. 3319–3328).
[7] Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296.
[8] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
[9] Molnar, C. (2018). Interpretable Machine Learning. Lulu.com.
[10] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), 93:1–93:42.