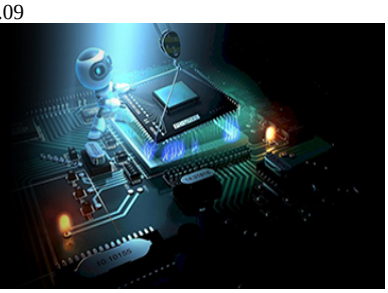
International Journal of Engineering in Computer Science



E-ISSN: 2663-3590
P-ISSN: 2663-3582
Impact Factor (RJIF): 5.52
www.computersciencejournals.com/ijecs
IJECS 2026; 8(2): 01.09
Received: 04-02-2026
Accepted: 06-03-2026

Dr. Punit Kumar Chaubey
Associate Professor, Department of Computer Science and Engineering (AI), Bansal Institute of Engineering and Technology, Lucknow, Uttar Pradesh, India

Dr. Sunita
Professor, Department of Computer Science, SAITM, Gurugram, Haryana, India

Rupendra Kumar
Assistant Professor, Department of Computer Science, GNIOT Institute of Professional Studies, Greater Noida, Uttar Pradesh, India

Gurpreet Kaur
Assistant Professor, Department of Computer Applications, Hierank Business School, Noida, Uttar Pradesh, India

Vibhanshu Tripathi
Assistant Professor, BCA/MCA/B.Tech, Institute of Technology and Management, GIDA, Gorakhpur, Uttar Pradesh, India

Corresponding Author:
Dr. Punit Kumar Chaubey
Associate Professor, Department of Computer Science and Engineering (AI), Bansal Institute of Engineering and Technology, Lucknow, Uttar Pradesh, India

Explainable AI (XAI): Bridging the gap between black-box models and human trust

Punit Kumar Chaubey, Sunita, Rupendra Kumar, Gurpreet Kaur and Vibhanshu Tripathi

DOI: https://www.doi.org/10.33545/26633582.2026.v8.i2a.270

Abstract
Arguably, the most recent years have been characterized by unprecedented growth of the concept of Artificial Intelligence (AI) with the popularization of massive machine learning models and deep learning models. Despite the excellent predictive power of these models, transparency, accountability and trust have been raised due to the fact that these models are opaque. The result of this has led to the creation of Explainable Artificial Intelligence (XAI), a paradigm that seeks to offer AI systems meaning and comprehensibility to humans. The article is a review of theoretical foundations of the XAI, its techniques, and uses, and the significance of the concept in the environment of bridging the trust-black-box gap between human trust and black-box models. The paper gives an in-depth taxonomy of XAI tools, such as intrinsic interpretability and post-hoc explanations, and some of the most prominent methods, such as LIME, SHAP, Grad-CAM, and counterfactual explanations. In addition, it also discusses such matters as model fidelity, robustness and ethical issues. XAI contribution to such significant spheres as healthcare, financial, and security is also addressed. Finally, the paper offers the future research directions, which are hybrid models, regulatory frameworks, and human-centered AI design. The findings point to the importance of XAI in ensuring transparency, equity, and reliability of the modern AI systems.

Keywords: Explainable AI, Interpretability, Black-box models, LIME, SHAP, Trust in AI, Transparency, Machine Learning

1. Introduction
Artificial Intelligence (AI) is a disruptive technology that has changed many spheres of life and performed an essential influence on such areas as healthcare, finance, transportation, and cybersecurity. The recent breakthrough of deep learning and advanced machine learning algorithms has helped systems to handle large volumes of information and provide highly precise predictions and automated decision making systems. Such models especially the deep neural networks have shown to be better in image recognition, natural language processing and predictive analytics amongst other things. Nevertheless, most of these systems are rather efficient yet they are black-box systems, and the inner mechanisms as well as decision-making paths are not visible to users and stakeholders [1]. This uninterpretability does not mean that human beings can comprehend, believe and certify the results of such systems.
Black-box models are not very transparent which poses a serious problem particularly in high stakes where decisions made are critical. The AI systems are getting more and more utilized in spheres like healthcare to help with the diagnosis of the disease, planning of treatment, and monitoring of the patient. Under these conditions, clinicians should be capable of seeing the rationale behind the AI-generated recommendations to guarantee safety and ethical accountability of patients. On the same note, automated systems are applied in credit scoring, fraud detection, and investment decision making especially in the financial sector. The systems must have transparency so that the regulatory bodies can be assured of fairness, elimination of discrimination and accountability. Lack of the ability to comprehend the AI decisions may result in the mistrust, possible biases, and ethical issues, thus deter the usage of AI technologies in the sensitive applications [2].
The higher application of AI has in its turn led to the amplification of the necessity to keep transparency, accountability, and fairness in automated decision making systems. Algorithms have been a cause of significant ethical and legal concerns because of their bias,

inexplicability, and incomprehensibility. Indicatively, skewed training information may lead to discrimination, and lack of clear explanations means that it may be hard to detect and rectify such problems. Along with that, explainability is a new area of legislation that is also a good sign of that there is still a need to work on AI systems that will be capable of giving explanations to their actions in a manner that will be understandable by humans. The problems make it clear that one should not limit himself to the models of pure performance but to systems that are interpretable, as well as trustworthy.

Explainable Artificial Intelligence (XAI) is a potential way out of these problems and it assists in enhancing the transparency and interpretability of AI systems. XAI is interested in developing approaches that enable users to understand how models come up with certain predictions and therefore, making complex systems more transparent and accountable. The XAI gives the meaningful explanations which allows the user to know variables influencing the choice of the model and to work on his or her future errors or biases and improve the system reliability. Such methods as feature importance analysis, local explanation model and visualization techniques have been embraced in order to analyze complex models [3]. These techniques do not only boost the confidence of the user but also make the interaction between man and the artificial intelligence system to be effective.

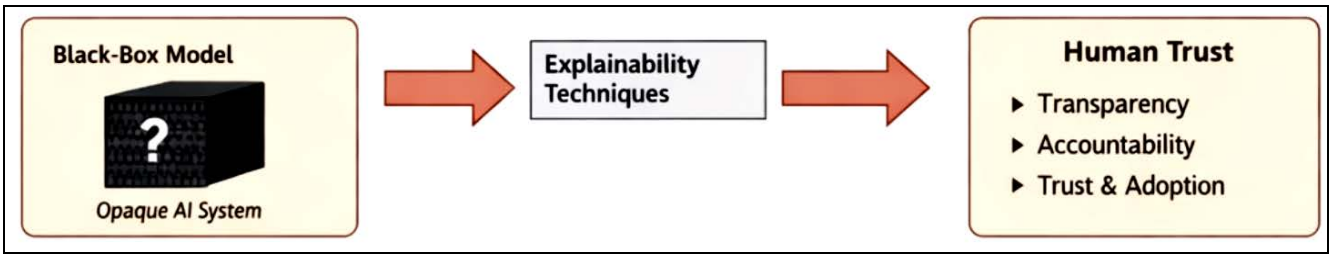


Fig 1: Conceptual Framework of Explainable Artificial Intelligence

XAI is also essential to develop human trust in AI technologies in addition to enhancing transparency. AI implementation in real-life applications is never possible without trust, especially in a situation where human lives and the overall outcomes of society are in question. The users can more easily accept and trust AI-driven decisions when they have clear and understandable descriptions. In addition, explainability is instrumental in supporting accountability because it allows the stakeholders to audit and analyze how the system behaves. Consequently, the XAI will help to establish ethical, fair, and responsible AI systems.

In this paper, the review of Explainable Artificial Intelligence with its main methodologies, practical usage, and the challenges will be provided. It looks at the different techniques of interpretability and how they have been effective in closing the gap between the complex black-box models and understanding by human beings. Moreover, the paper addresses the drawbacks of the existing methods as the problems of scalability, robustness, and the measures of evaluation. Lastly, it also discusses the future research directions that will be taken to improve explainability, introduce human-centered design, and create a set of standardized frameworks that ensure trustworthy AI. The key aim of this paper is to examine how XAI would be useful in mediating the difference between black-box models and human trust to permit the responsible and transparent implementation of the AI systems in various fields.

2. Background and Motivation

A. Black-Box Models in AI

The current Artificial Intelligence (AI) systems, especially the ones relying on deep learning design, are defined by a high degree of complexity and non-linearity. Such models include deep neural networks (DNNs), convolutional neural networks (CNNs), and transformer-based models and have a series of layers and can have millions or even billions of parameters. Although this complexity allows them to acquire complex patterns on a large-scale data and perform at the state-of-the-art on the different tasks, it also complicates their internal decision-making process, making it hard to interpret [4]. The inputs and outputs relations are coded in a non-transparent and distributed fashion, thus making it impossible to the user to grasp how certain predictions are made with ease.

Such a black-box approach of AI models has severe constraints, particularly in the areas that are most vital like healthcare, finance, and the legal system. In such situations, the results of decisions are usually grave, and the stakeholders demand the reasons of automated results to be clear. Indicatively, an AI-generated medical diagnosis should be explainable to clinicians so that patients are safe and to gain trust in the system. Equally, in financial decision-making, there is a need that the institutions can provide an explanation to either credit approvals or credit rejections as required by the regulations. Their inability to provide transparency with black-box models does not allow their use in this high-stakes setting even though they may be better predictors. This leads to the increased understanding that accuracy is not enough and interpretability needs to be incorporated in the design of the AI systems.

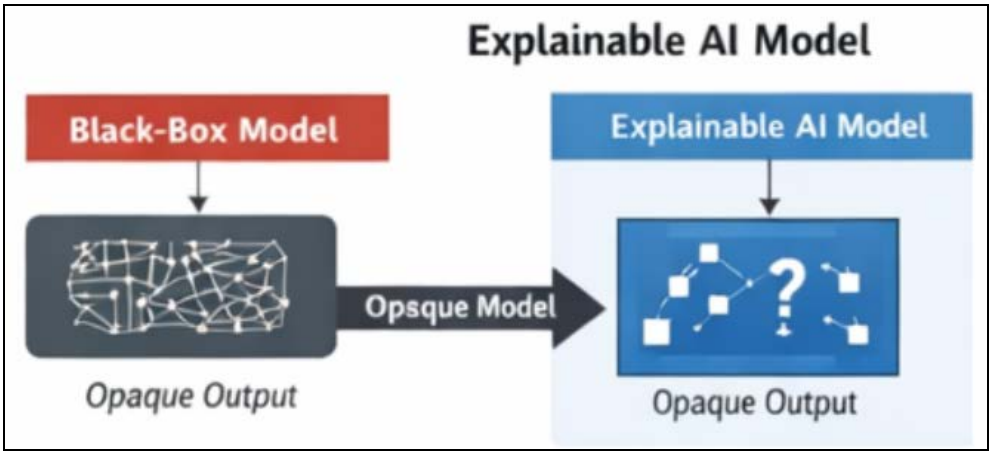


Fig 2: Architecture of Black-Box vs Explainable AI Models

B. Requirement to be Explainable

The growing use of AI systems has brought to the fore the urgent need to have explainability, due to various factors, which are closely interconnected. To begin with, user acceptance is based on trust and transparency. The more users comprehend the process of making decisions by AI systems and can confirm the rationale of their decisions, the more likely they are to use them [5]. Even very precise systems without transparency are going to be perceived skeptically.

Second, explainable AI has now been driven by regulatory compliance. The legal frameworks like the General Data Protection Regulation (GDPR) focus on the right to explanation as organizations must offer any meaningful information on automated decision-making processes [14]. This has forced industries to use AI systems which are capable of providing insights which are interpretable.

Third, explainability is critical in detecting errors and validation of a model. Through studying the way a model comes up with its predictions, developers and stakeholders are able to find out the biases, inconsistencies and possible weaknesses of the system [8]. It is especially essential in avoiding undesirable discriminatory results, which can occur because of biased training data or incorrect model design.

Lastly, the issue of ethics also supports the explainability even more. AI systems should be fair, accountable and transparent in their operation as a way of ensuring that they conform to values and ethical principles of the society [12]. Explainable AI ensures that interested parties can determine the decisions taken to be impartial and explainable, hence, responsible AI implementation.

C. Interpretability vs Explainability

Interpretability and explainability are two concepts in AI that are similar but different as they are used interchangeably. Interpretability is the degree to which a human being can directly comprehend the inner processes and the inner structure of a model. Interpretable models, like the linear regression or decision trees, give easy information on how the inputs are translated into outputs.

Conversely, explainability is concerned with whether or not a model, especially black-box models, can give intelligible explanations of its predictions, but not necessarily its internal processes [11]. The explicability is normally obtained by use of post-hoc methods that approximate or describe the model behavior in a manner that is understandable by human beings.

It is important to know the difference between the two concepts in order to create effective AI systems. Whereas interpretability focuses on simplicity and transparency, explainability can also use complex models and still be able to extend some meaningful insights on the decision-making processes of the model. The combination of the two forms the basis of Explainable Artificial Intelligence (XAI), and under these conditions, it is possible to create powerful and trustworthy systems.

3. Taxonomy of Explainable AI

A. Intrinsic Interpretability

Intrinsic interpretability is the property of some machine learning models which are per se transparent and easy to interpret because of their simple structure and mathematical formulation. These models are constructed in a manner where the internal logic can be directly interpreted with no need of the use of any other method of explaining the logic. Well known models that are intrinsically interpretable are linear regression, decision trees and logistic regression. In linear regression, the input features are directly correlated with the output as the relationship is clearly defined by coefficients, and one can easily see the effect of each feature. On the same note, decisions made using decision trees are clear in the sense of being structured in a rule-based manner where a decision is arrived at as a result of a sequence of interpretable conditions. Logistic regression which is commonly applied in the classification task also provides interpretability with feature weights which reflect the effect of variables on the outcome that is predicted [24].

The main benefit of intrinsic interpretability is that this allows transparency and trust without incurring extra computation cost. The models come in handy especially where explainability is a requirement like in the fields of healthcare and finance. But as they are simple they tend to miss non-linear relations in data which is often complex leading to poor predictive behavior of the data as compared to other advanced black box models. Thus, intrinsically interpretable models are easy to understand, but such models are not always applicable in very complex tasks.

B. Post-hoc Explainability

The post-hoc explainable techniques are created to describe opaque and complex models once they are trained. These methods do not involve adjustments to the underlying model, as opposed to intrinsically interpretable models, but

are instead meant to give explanations of the predictions of the underlying model. The post-hoc techniques are also needed to comprehend black-box models like deep neural networks where it is impossible to interpret them directly.

A number of forms of post-hoc explanation methods have been put forward. The importance of features methods determine the most important input features to model predictions. These techniques assist the end-users to know the relative significance of variables in making decisions. Alternatively, surrogate models may be used which entails training a more interpretable model to predict the behavior of the complex model. This enables the users to have a glimpse of the general operation of the system. Also, visualisation, including saliency maps and heatmaps, is a popular deep learning technique that can be used to identify relevant areas in input data, especially in image and text-based systems [9].

The methods of post-hoc are flexible and may be applied to very diverse models. At times, however, they can give rough explanations which do not well represent the actual internal mechanism of the original model, which casts doubt on the fidelity and reliability.

C. Model-Specific and Model-Agnostic Approaches

The applicability of explainability techniques can also be defined as per the type of models in which they can be applied. Model-specific approaches are model-specific, i.e. targeted at specific model architectures and they use the internal structure of the model to produce explanations. As an illustration, Grad-CAM is designed to work with convolutional neural networks (CNNs) and relies on the information on the gradient to generate visual explanations of specific important areas in images [17]. These approaches have a tendency to give extremely precise and detailed explanations but are only applicable to a kind of model.

By comparison, model-agnostic algorithms are model-agnostic and may be used on any machine learning algorithm. Some of the techniques that fall under this category are LIME and SHAP, which tries to model the behavior of a model by approximating the behavior using interpretable representations and there is no need to access the internal model parameters [3], [4]. Model-agnostic methods are more flexible and are extensively applied in practice, but they can also require more computation.

D. Local vs Global Interpretability

The other difference in XAI is that of local and global interpretability. Local explanations are aimed at explaining individual predictions in terms of the impact of various input characteristics on a certain output. Case-by-case analysis is applicable to these explanations and they are typically applied in medical diagnosis and fraud detection.

On the other hand, the purpose of global interpretability is to be able to introduce a general idea of how the model behaves in the whole dataset. This involves the determination of broad trends, importance of features, and limits of decisions that determine predictions of the model [13]. Global explanations are useful in the validation of the model, debugging and the consistency in decision making.

Explainable AI has both a local and a global interpretation, which have a complementary role. Whereas local explanations give a detailed explanation of certain instances, global ones give a larger picture of model behavior, which is a part of the overall picture of AI systems.

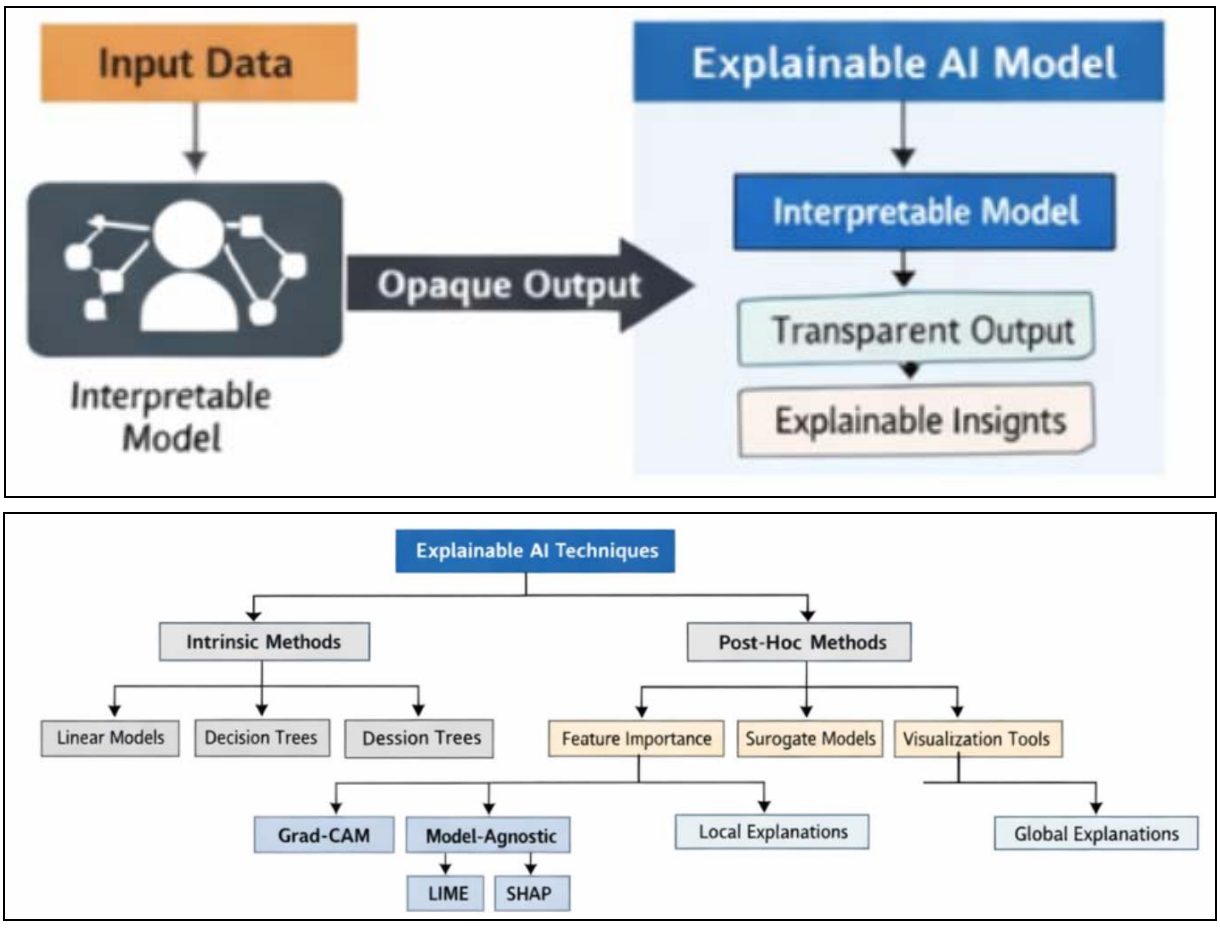


Figure 3: Taxonomy of Explainable AI Techniques

4. Key XAI Techniques

A. LIME (Local Interpretable Model-Agnostic Explanations)

LIME can be considered one of the most popular model-agnostic explanation methods that aim at local interpretation of complex machine learning models. The basic principle of LIME is to estimate the behavior of a black-box model in the locality of a particular prediction by a more understandable model (e.g. a linear model or a decision tree). In order to do this, LIME perturbs the input data and creates synthetic samples around the instance of interest and monitors the changes in the model predictions in response to these perturbations. According to this analysis, features are given importance weights by LIME, which shows that they make contributions towards prediction [3].

The flexibility of LIME is also one of its most important advantages since it can be used to apply it to any model, such as deep neural networks, ensemble models, and support vector machines. It also offers user friendly explanations that one can comprehend without difficulties. Nonetheless, LIME explanations can only be explained locally and might be different with respect to the sampling procedure that may compromise stability and reliability.

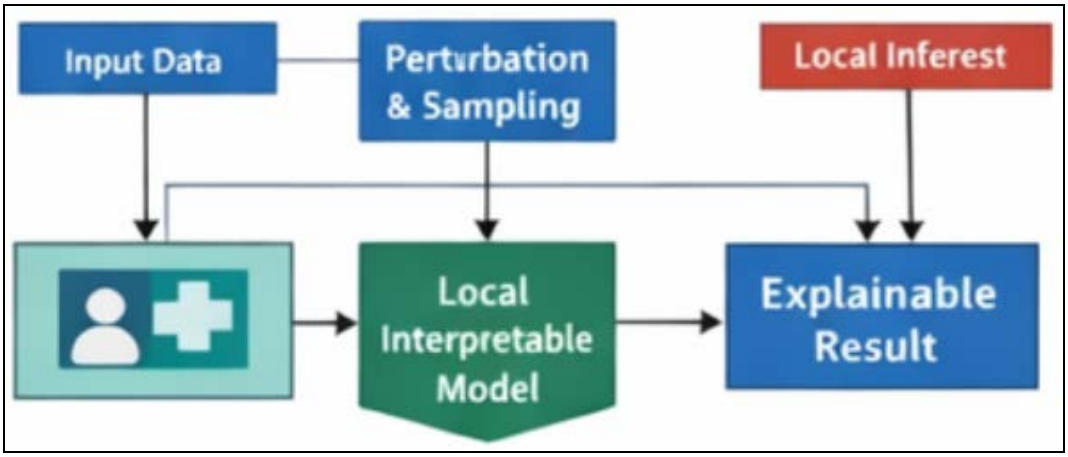


Fig 4: Workflow of LIME Explanation Process

B. SHAP (SHapley Additive exPlanations)

SHAP is an explanation procedure that is developed on the basis of Shapley values of cooperative game theory. It gives a value of importance to every feature based on the contribution of the feature to the prediction when all the combinations of features are possible. The latter guarantees consistency of the explanations, their addition and fairness, and fulfills major requirements like local accuracy and consistency [4].

SHAP has local and global interpretability. In the local case, it attributes individual predictions with the help of quantifying the contribution of each feature. It is applicable globally to examine the importance of features in general to the dataset. The good mathematical basis of SHAP is one of the strengths of the method because it makes it more dependable than the approaches based on heuristics. Nevertheless, the computation of the precise Shapley values may be computationally intensive, particularly when one has large datasets and complicated models. This notwithstanding, SHAP has emerged as a conventional tool of XAI because of its strength and interpretability.

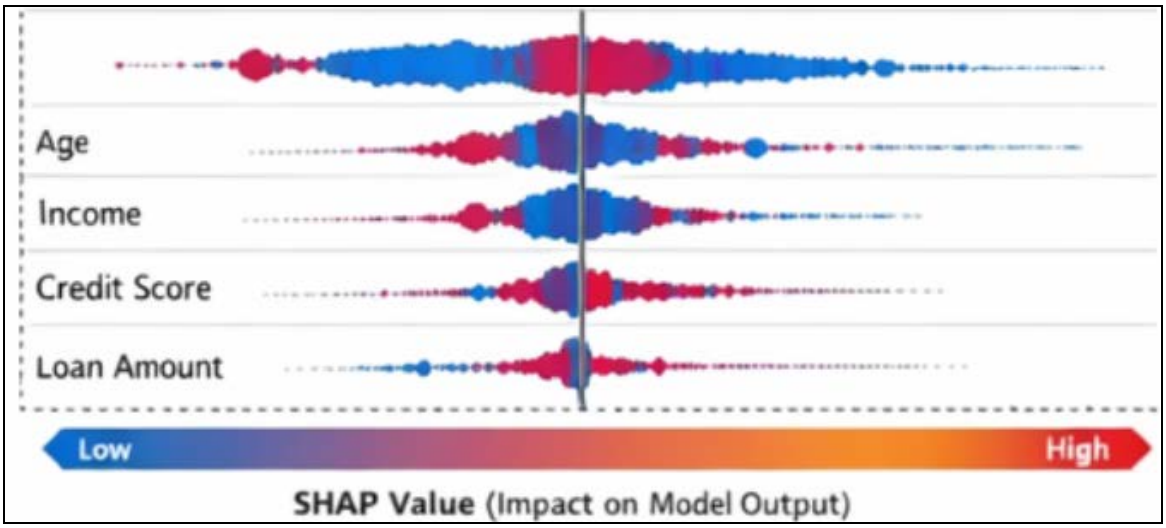


Fig 5: SHAP Value Interpretation Graph

C. Grad-CAM (Gradient-weighted Class Activation Mapping)

Grad-CAM is a model-specific explainability method that can be used with convolutional neural networks (CNNs) when applied to computer vision. It produces visual explanations with the help of gradient information that

flows into the last convolutional layers of a network. The heatmaps are created using these gradients to indicate the parts of an input image that have the most impact on the prediction of a model [17].
Grad-CAM will particularly be applicable in scenarios that require visual interpretability, including medical imaging, object detection, and facial recognition. It also shows critical areas in images, which makes the user know what the model is doing in its decision making. Grad-CAM can however be used exclusively on CNN-based architectures and not on other models.

D. Counterfactual Explanations
Counterfactual explanations give us knowledge on how models behave, what the minimal modification of the input features would lead to a different prediction. That is, they respond to the question of what happens to an outcome when one makes small changes to the input data. To illustrate, in a loan giving system, a counterfactual explanation would give the indication that the same amount of income increment would alter the outcome of the decision between rejection and approval [14].
Such descriptions are very useful to the end-users, as they are practical and easy-to-understand. They assist the users to know the limits of the decisions and give directions to them on how to attain what they want. Also, the counterfactual explanations do not involve the inside structure of the model and hence can be applied to black-box systems. Nonetheless, it may not be easy to produce realistic and meaningful counterfactuals, particularly when dealing with high-dimensional space of data.

E. Layer- wise Relevance Propagation (LRP)
The Layer-wise Relevance Propagation (LRP) is a model that is applicable in interpreting deep neural networks, by breaking down the prediction to the input features of the network. This is done by backward propagation of the prediction score the network layers, and the relevance of each neuron is apportioned to its parents. The output of this process is a relevance score of each input feature that shows how much the input feature contributes to the ultimate prediction [22].
LRP is effective especially in the analysis of deep learning models applied in text and image processing. It gives detailed information on the effect of various elements of the input on the model output and is thus useful in debugging the model as well as enhancing the performance of the model. Similarly to Grad-CAM, LRP is also model-specific and needs the internal architecture of the network.
On the whole, these XAI methods are important in increasing the extent of transparency and interpretability of AI systems. They can address this issue by offering substantive explanations to support the creation of a gap between complex black-box models and human knowledge, therefore, creating trust and responsibility in AI applications.

5. Applications of XAI

A. Healthcare
Explainable Artificial Intelligence (XAI) is becoming increasingly relevant to the healthcare sector, with the outcomes of the decisions having a direct impact on the patient safety and welfare. The most typical AI models are regarded to be the disease diagnosis, medical analysis of images, drug discovery, and prediction of patient risks. However, the adoption of these systems is extremely cumbersome regarding trust amongst the clinicians and medical practitioners. XAI can also satisfy this need to provide interpretable model predictions, so medical professionals are aware of the functioning of automated decision-making. One such method is the Grad-CAM method that produces visual heatmaps that locate some regions in a medical image, e.g. a tumor or lesion, in the diagnosis of the model [38]. This is not only enhancing transparency, but will also assist clinicians to check on AI generated results. Besides, explainability is applied to eliminate biases in medical data, which is critical to making just and reasoned treatment recommendations. Therefore, XAI is significant to improve the trustworthiness, responsibility, and acceptability of AI systems in medical practice.

B. Finance
Financial industries have been employing AI in credit scoring systems, fraud detection systems, algorithmic trading systems, as well as risk management systems. These uses entail life and death decisions which may have severe consequences on people and organizations. Thus, accountability and transparency are important requirements. XAI allows financial institutions to explain and understand automated decisions, which also increases the level of trust in the user and regulators. As an example, with credit scoring system, the explainability techniques can be used to determine the factors that can lead a loan to be approved or rejected and the institutions can clearly explain the reasons to their customers. Moreover, in fraud detection systems, XAI is used to assist the analysts to see how AI models identified a pattern or anomaly in their data, which makes the decision making and investigation processes easier. The regulatory frameworks also highlight the importance of explainability since organizations should be guided by the law on the aspects of fairness and non-discrimination [12]. XAI makes financial AI systems transparent and ethical, as well as adhering to regulatory requirements by offering interpretable insights.

C. Cybersecurity
Another area in which the XAI has been of great importance is through cybersecurity. The AI-based systems are extensively employed so that to identify cyber threats such as malware, phishing attacks, and network intrusions. Although these systems are very useful in the detection of suspicious activities, they are black-box in nature, which may render the security analysts unable to know how the alerts and classifications were reached. XAI in this regard solves this problem by offering a description of model forecasts so that analysts can understand and justify the selections of AI systems. As an illustration, explainable models can showcase certain attributes or trends in network traffic that denote malicious activities to enhance situational awareness and response plans [37]. Moreover, XAI will help to increase the confidence in automated cybersecurity systems and minimize false positives and improve threat analysis. This results in the enhanced reliability of the system and the increased security measures against the emergent cyber threats.

D. Autonomous Systems
Self-driving cars and intelligent robotics are all examples of

autonomous systems that make important decisions in real-time using AI. These are systems that work in dynamic and unforeseeable conditions where safety and accountability are the most important. XAI is a significant factor in making sure that all the decisions of autonomous systems are comprehensible. As an example, in autonomous cars, XAI may be used to give an explanation of certain behavior, such as braking, switching lanes, or avoiding obstacles, to developers and users to know how the system works. This is mainly essential during the instances of accidents or failure of the system that would demand accountability and traceability. Besides, explainability helps to debug and improve autonomous systems through the detection of possible errors or biases during decision-making. XAI will help to promote the safe and responsible use of autonomous technologies by promoting their transparency and trust.

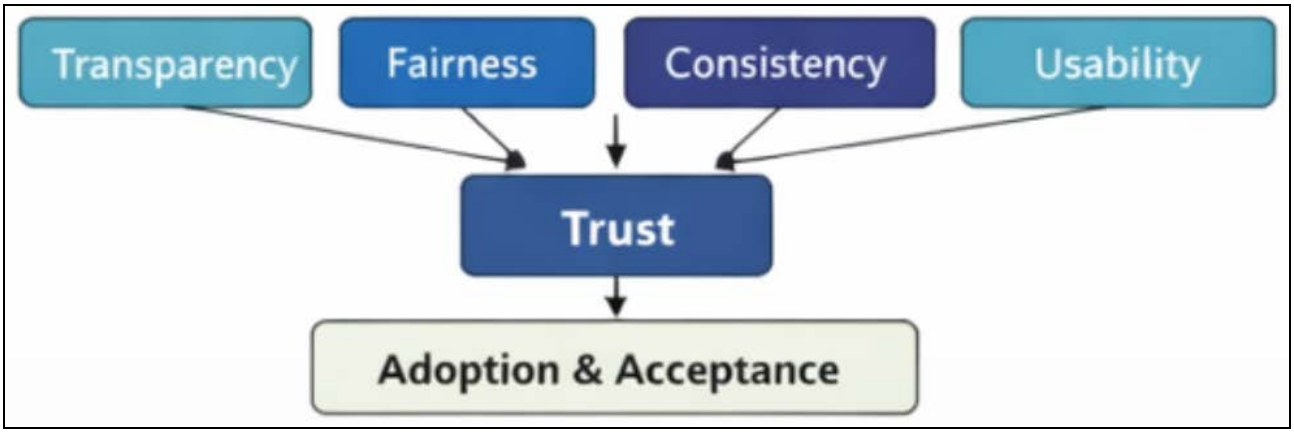


Fig 6: Application Areas of XAI across Domains

6. Evaluation of XAI

The effectiveness of the Explainable Artificial Intelligence (XAI) techniques is a complicated problem to evaluate, and the main reason for this is the subjectivity of the interpretability. There are a number of measures of evaluation that are frequently employed to evaluate XAI methods. The degree to which the explanations are true to the actual behavior of the underlying model is represented by fidelity. The process of decision-making of the AI system can be accurately explained in a high-fidelity form. Interpretability evaluates the ease at which human beings can interpret the explanations made by the model. A more explainable version is important to make sure that the end-users are able to make meaningful conclusions out of the system outputs. Consistency measures the permanence of explanations on similar inputs such that the process of explanation yields similar outputs when the similar data is given. Completeness is a parameter that determines how well an explanation is able to describe the complete behavior of the model; all the input features and interactions involved [13].

In addition to these quantitative measures, human-based evaluation techniques are required in the process of establishing the utility of the explanations in practice. Such methods take into account the context within which the AI system is going to be used and make the explanations comprehensible, practical, and relevant in the cultural context of the target users. XAI is a concept that needs technical rigor, as well as knowledge about the way people interact with and perceive AI systems.

7. Challenges

Although there has been tremendous improvement, XAI has various challenges that it has been struggling with. Trade-off between the accuracy and interpretability is one of the most basic problems. Decision trees or linear regression are such very interpretable models that are often less predictive than more complex, non-interpretable models, such as the deep neural networks [8]. The balance between the complexity of the models and their interpretability is one of the primary problems that should be found when the successful XAI systems are being implemented.

The second challenge is that there are no generally accepted models of assessments which would enable the comparison of different XAI methods. Without a general measure, one can barely be aware of which of the techniques is more applicable in a particular application or even how the different approaches can be integrated with one another to operate in real-world systems. Scalability is also a problem especially in big scale and complex models. The algorithmics of the explanation generation task of such models can be prohibitive, limiting the scale of application of certain XAI techniques to high performance environments. Furthermore, some of the methods of the explanation may as well yield misdirected or simplified interpretations thereby, providing erroneous conclusions or causing the lack of trust in the system. Lastly, another reason why the explanation can be effective is the domain knowledge and cognitive abilities of the user which also leads to the human-centered design of XAI systems.

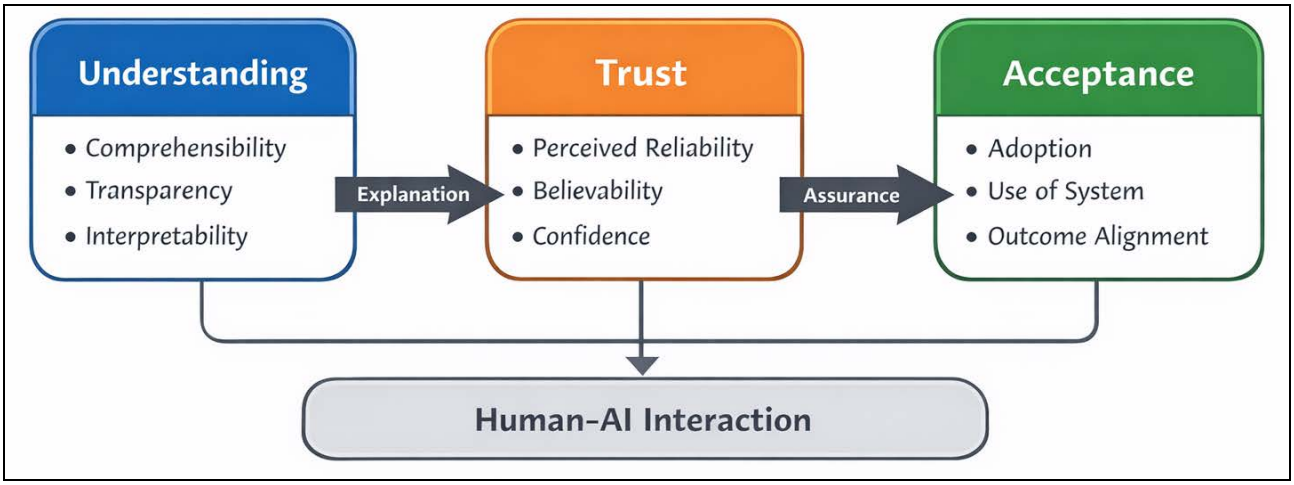


Fig 7: Human Trust Model in Explainable AI

8. Future Directions

The next significant domain of research on XAI is aimed at the creation of more reliable, interpretable and user-friendly methods. Human-centered XAI is focused on generating explanations that are specific to the needs of various groups of users and consider their knowledge, experience, and cognitive style. The incorporation of the element of causal inference into XAI is also likely to improve the interpretability of models by giving a better understanding of the causal relationship behind the models but not by merely correlative explanations [12]. A second direction is real-time explainability, in particular of dynamic and autonomous systems like self-driving cars, in which explainability should be availed on demand, as well as rapidly and consistently in reaction to real-time choices. In addition, the ethical AI theories and regulatory frameworks will be essential in informing responsible AI systems implementation and making sure that explainability will be in line with legal and ethical frameworks. With the advances in the field, XAI is likely to become the part of trustful, transparent, and responsible AI systems.

References

1. Adadi A, Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access. 2018;6:52138–52160. doi:10.1109/ACCESS.2018.2870052
2. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A review of machine learning interpretability methods. Entropy. 2021;23(1). doi:10.3390/e23010018
3. Ribeiro MT, Singh S, Guestrin C. Why should I trust you? Explaining the predictions of any classifier. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. doi:10.1145/2939672.2939778
4. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems. 2017. doi:10.48550/arXiv.1705.07874
5. Molnar C. Interpretable machine learning. 2019. doi:10.48550/arXiv.1909.09436
6. Gunning D. Explainable artificial intelligence (XAI). DARPA Program. 2017. doi:10.48550/arXiv.1901.09423
7. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. 2017. doi:10.48550/arXiv.1702.08608
8. Rudin C. Stop explaining black box models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 2019. doi:10.1038/s42256-019-0048-x
9. Montavon G, Samek W, Müller KR. Methods for interpreting and understanding deep neural networks. Digital Signal Processing. 2018. doi:10.1016/j.dsp.2017.10.011
10. Samek W, Wiegand T, Müller KR. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. ITU Journal. 2017. doi:10.48550/arXiv.1708.08296
11. Lipton ZC. The mythos of model interpretability. Queue. 2018. doi:10.1145/3236386.3241340
12. Mittelstadt B, et al. The ethics of algorithms: Mapping the debate. Big Data & Society. 2016. doi:10.1177/2053951716679679
13. Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence. 2019. doi:10.1016/j.artint.2018.07.007
14. Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box. Harvard Journal of Law & Technology. 2018. doi:10.48550/arXiv.1711.00399
15. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. Proceedings of the International Conference on Machine Learning. 2017. doi:10.48550/arXiv.1703.01365
16. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models. 2013. doi:10.48550/arXiv.1312.6034
17. Selvaraju RR, et al. Grad-CAM: Visual explanations from deep networks. Proceedings of the IEEE International Conference on Computer Vision. 2017. doi:10.1109/ICCV.2017.74
18. Ribeiro MT, Singh S, Guestrin C. Anchors: High-precision model-agnostic explanations. Proceedings of the AAAI Conference on Artificial Intelligence. 2018. doi:10.48550/arXiv.1802.07623
19. Slack D, et al. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2020. doi:10.1145/3375627.3375830

20. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. Proceedings of the International Conference on Machine Learning. 2017. doi:10.48550/arXiv.1704.02685
21. Ribeiro MT, et al. Model-agnostic interpretability of machine learning. 2016. doi:10.48550/arXiv.1606.05386
22. Bach S, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLOS ONE. 2015. doi:10.1371/journal.pone.0130140
23. Alvarez-Melis D, Jaakkola TS. On the robustness of interpretability methods. Proceedings of ICML Workshop. 2018. doi:10.48550/arXiv.1806.08049
24. Guidotti R, et al. A survey of methods for explaining black box models. ACM Computing Surveys. 2018. doi:10.1145/3236009
25. Mersha M. Explainable artificial intelligence: A survey of needs and challenges. Neurocomputing. 2024. doi:10.1016/j.neucom.2024.127403
26. Mohamed A, et al. Explainable artificial intelligence: A systematic review of systematic reviews. Array. 2025. doi:10.1016/j.array.2025.100327
27. Nazim S, et al. Explainable AI using SHAP, LIME and Grad-CAM. PLOS ONE. 2025. doi:10.1371/journal.pone.0318542
28. Salih A, et al. A perspective on explainable AI methods: SHAP and LIME. Advanced Intelligent Systems. 2024. doi:10.1002/aisy.202400304
29. Liu H, et al. Using SHAP and LIME to explain ML models. Healthcare Analytics. 2025. doi:10.1002/hca2.10089
30. Sathyan S, et al. Interpretable AI for biomedical applications. Diagnostics. 2022. doi:10.3390/diagnostics13040678
31. Zhao X, et al. BayLIME: Bayesian local interpretable model-agnostic explanations. 2020. doi:10.48550/arXiv.2012.03058
32. Sai Ram Aditya P, Pal M. Local interpretable SHAP explanations. 2022. doi:10.48550/arXiv.2210.04533
33. Panda M, Mahanta S. Explainable AI for healthcare applications. 2023. doi:10.48550/arXiv.2311.05665
34. Givisis I, et al. Comparing SHAP and LIME methods. Electronics. 2025. doi:10.3390/electronics14234766
35. Ahmed S. Comparative analysis of LIME and SHAP. 2024. doi:10.13140/RG.2.2.18864.42
36. Panigrahi S, et al. Hybrid SHAP-LIME framework for XAI. 2026. doi:10.1080/10589759.2026.2644404
37. Nazim S, et al. Explainable AI in malware detection. 2025. doi:10.1016/j.cose.2025.103214
38. Chaddad A, et al. Explainable AI for healthcare: Applications and challenges. IEEE Reviews in Biomedical Engineering. 2021. doi:10.1109/RBME.2020.3030807
39. Guidotti R, et al. Explainable AI methods: A survey. Information Fusion. 2019. doi:10.1016/j.inffus.2018.10.011
40. Arrieta AB, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges. Information Fusion. 2020. doi:10.1016/j.inffus.2019.12.012