iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VI **Month of publication:** June 2026

DOI: https://doi.org/10.22214/ijraset.2026.83361

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 14 Issue V May 2026- Available at www.ijraset.com

Explainability in Modern Artificial Intelligence: A Review of Techniques, Applications, and Limitations

Nisha Rani[1], Dr. Gaurav Aggarwal[2]
[1]*Ph.D. Research Scholar, Computer Science & Engineering, Jagannath University, Bahadurgarh*
[2]*Associate Professor, Computer Science & Engineering, Jagannath University, Bahadurgar*

Abstract: ***As the use of deep learning and complex machine learning models, whose decision-making procedures are frequently opaque, increases, Explainable Artificial Intelligence (XAI) has become a crucial field of study. Even while these models are highly predictive, their opaque nature raises fundamental questions about accountability, transparency, justice, and trust, especially in high-stakes industries like healthcare, banking, law, and autonomous systems. This essay offers a thorough analysis of Explainable Artificial Intelligence, methodically examining its basic ideas, definitions, and taxonomy. We evaluate popular approaches including LIME, SHAP, Grad-CAM, Integrated Gradients, surrogate models, and counterfactual explanations and classify XAI strategies according to model dependencies, explanation processes, and explanation scope. Additionally, we examine real-world XAI applications in a variety of fields, emphasizing interpretability requirements unique to each domain. Important issues are rigorously evaluated, such as the trade-off between interpretability and accuracy, the absence of established evaluation measures, subjective human judgment, and ethical considerations.***

Finally, the report concludes by outlining new areas of research, including robustness of explanations, human-centered explainability, interactive and real-time explanations, and standardized evaluation frameworks. In order to help academics, practitioners, and policymakers create transparent, reliable, and responsible AI systems, this review is intended to be a useful resource.

Keywords: Explainable Artificial Intelligence (XAI), Interpretability, Transparency, Post-hoc Explanations, Human-Centered AI, Evaluation Metrics, Trustworthy AI, Deep Learning.

I. INTRODUCTION

Over the past ten years, artificial intelligence (AI) and deep learning approaches in particular have grown rapidly due to developments in learning algorithms, large-scale data availability, and processing capacity. In a variety of fields, such as computer vision, natural language processing, healthcare diagnostics, financial forecasting, autonomous systems, and decision-support tools, these models have shown impressive performance. Because of this, AI systems are being used more and more in real-world settings where their choices have a big influence on people and society. Despite their impressive performance, many state-of-the-art AI and deep learning models operate as black-box systems, meaning their internal decision-making processes are not easily interpretable by humans. Deep neural networks and other complex designs frequently lack transparency, which makes it challenging to comprehend how input attributes affect outputs or the reasoning behind a given prediction or choice. Particularly when AI systems are employed in situations requiring essential decision-making, this opacity poses grave questions about trust, accountability, fairness, and dependability. In high-stakes fields like healthcare, finance, law, autonomous driving, and public policy, explainability becomes especially important. Misdiagnosis, financial loss, legal injustice, and threats to human safety are just a few of the serious outcomes that can result from making biased or incorrect decisions in these areas. AI systems are increasingly required by ethical standards and regulatory frameworks to give clear and convincing justifications for their choices. It is difficult to verify model behavior, identify biases, guarantee regulatory compliance, and win user trust without explainability.

In response to these difficulties, Explainable Artificial Intelligence (XAI) has become a crucial field of study with the goal of creating strategies and tactics that improve the transparency, interpretability, and reliability of AI models without appreciably sacrificing performance. Model-specific, model-agnostic, local, and global explanation strategies are just a few of the many XAI approaches that have been put out in recent years. However, it has been challenging for scholars and practitioners to develop a shared understanding of the discipline due to the variety of approaches, assessment criteria, and application scenarios.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

A. Motivation for this review

The necessity for a thorough review that methodically examines current explainability methodologies, their advantages and disadvantages, and their cross-domain applicability is evident given the quick growth of XAI research and its increasing significance in practical applications. By assisting academics, developers, and policymakers in choosing suitable XAI techniques, such a review can help close the gap between theoretical advancements and real-world implementation.

B. Contributions of this Paper Include

- A comprehensive overview of Explainable Artificial Intelligence concepts and terminology
- A structured taxonomy of XAI techniques based on model type, explanation scope, and methodology
- A comparative analysis of popular XAI methods, highlighting their advantages and limitations
- A discussion of evaluation strategies and challenges in measuring explainability
- An examination of XAI applications in high-stakes domains and open research challenges.

II. BACKGROUND AND FUNDAMENTAL CONCEPTS

The literature uses a range of overlapping and occasionally conflicting names to express explainability-related ideas as Explainable Artificial Intelligence (XAI) continues to develop. Ambiguity in research and real-world applications may result from this absence of defined nomenclature. This section provides a clear and consistent overview of the core ideas of XAI, focusing on the differences between explainability and interpretability, the connection between accountability, transparency, and trust, and the distinctions between interpretable and black-box models.

A. Explainability vs. Interpretability

Although interpretability and explainability are closely related ideas, they are not the same and are frequently used interchangeably in the literature. The term "interpretability" describes a model's innate capacity for direct human understanding. An interpretable model eliminates the need for extra explanation methods by enabling users to understand how inputs are converted into outputs. Rule-based systems, decision trees, and linear regression are a few examples of systems with clear and traceable decision logic.

Explainability, on the other hand, concentrates on offering post hoc justifications for a model's actions, especially in situations when the model is not naturally interpretable. The goal of explainability techniques is to provide a human-understandable summary or approximation of the internal logic of complex models. These explanations could be global, providing insights into the general behavior of the model, or local, outlining specific expectations. Because direct interpretation is not possible for sophisticated architectures like deep neural networks, explainability is particularly important.

In summary, interpretability is a property of the model itself, while explainability is a characteristic of the tools and methods applied to describe model behavior. This distinction is crucial for evaluating XAI approaches and selecting appropriate methods based on application requirements.

Table 1. Explainability vs. Interpretability in AI

Aspect	Interpretability	Explainability
Definition	Degree to which a model's internal logic can be directly understood by humans	Ability to provide understandable reasons for a model's predictions
Nature	Inherent property of the model	Achieved using additional techniques or methods
Applicability	Mostly applies to simple or transparent models	Primarily used for complex or black-box models
Examples	Linear regression, decision trees, rule-based models	LIME, SHAP, saliency maps, counterfactual explanations
Scope	Global understanding of model behavior	Can be local (single prediction) or global
Dependency	Model-dependent	Can be model-agnostic or model-specific

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

B. Transparency, Trust, and Accountability in AI

- Transparency: Transparency in AI refers to the extent to which the processes, assumptions, and decision-making mechanisms of an AI system are open and understandable to stakeholders. Users, developers, and regulators can examine how decisions are made, what data is used, and what influences results with a transparent system. A fundamental prerequisite for meaningful explainability is transparency.
- Trust: Trust in AI systems emerges when users believe that the system operates reliably, fairly, and consistently within its intended scope. By enabling users to evaluate the reasoning behind AI-generated judgments, explainability plays a crucial role in fostering this trust. Without clear explanations, users may be hesitant to rely on AI systems, especially in domains that are safety-critical or ethically sensitive.
- Accountability: Accountability relates to the ability to assign responsibility for AI-driven decisions and outcomes. In the absence of explainability, determining whether a system behaved appropriately—or identifying the source of an error—becomes extremely difficult. By facilitating audits, compliance checks, and the detection of biases or inadvertent behaviors, transparent and explainable AI systems promote accountability. Transparency, trust, and accountability work together to create a closely linked framework that supports the appropriate application of AI.

Table 2. Key Concepts Supporting Responsible AI

Concept	Description	Role in XAI
Transparency	Openness of model structure, data usage, and decision logic	Enables inspection and understanding
Trust	User confidence in system reliability and fairness	Encourages adoption and reliance
Accountability	Ability to assign responsibility for AI outcomes	Supports audits and legal compliance

C. Black-Box vs. Interpretable Models

Depending on how visible their underlying decision-making processes are, AI models can be broadly classified as either interpretable or black-box. Complex internal structures that don't offer clear explanations for their predictions are a hallmark of black-box models. This category usually includes large-scale machine learning models, ensemble approaches, and deep neural networks. Although these models frequently perform well in terms of prediction, their opacity restricts comprehension and examination.

Conversely, interpretable models are made to be comprehensible by people without the need for additional explanation methods. Users can follow the explicit decision logic in these models to see how certain inputs result in particular outputs. Interpretable models have benefits in terms of transparency, debugging, and regulatory compliance, but they may lose some predicted accuracy in extremely complicated jobs.

In AI research, the trade-off between interpretability and accuracy has long been a major concern. By creating techniques that improve the explainability of black-box models while maintaining their performance advantages, XAI aims to close this gap. Selecting appropriate AI solutions requires an understanding of the differences between different model types, especially in high-stakes situations where explainability is a crucial necessity.

Table 3. Comparison of Black-Box and Interpretable AI Models

Feature	Black-Box Models	Interpretable Models
Internal structure	Highly complex and opaque	Simple and transparent
Human understanding	Limited or unavailable	Directly understandable
Predictive performance	Typically high	Moderate to high
Debugging difficulty	Difficult	Easier
Regulatory compliance	Challenging	Easier to justify
Common examples	Deep neural networks, ensemble models	Decision trees, linear models

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

III. TAXONOMY OF EXPLAINABLE AI TECHNIQUES

Explainable Artificial Intelligence (XAI) methods can be systematically categorized based on model dependency, scope of explanation, and explanation mechanism. This taxonomy provides a unified framework to understand, compare, and select appropriate explanation techniques for different application scenarios.

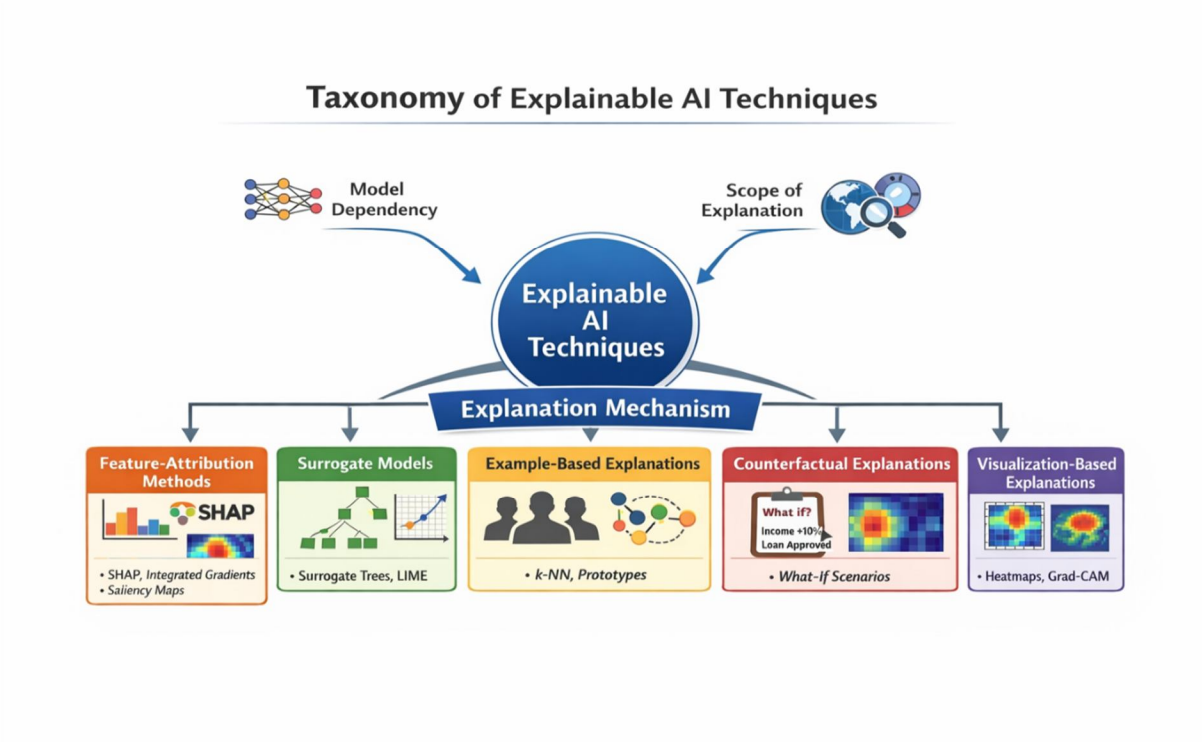


Fig 1. Taxonomy of XAI Techniques

This category classifies XAI techniques based on *how* explanations are generated.

1) Feature-Attribution Methods

These methods assign **importance scores** to input features based on their contribution to a prediction.

Examples:

- SHAP
- Integrated Gradients
- Feature importance scores
- Saliency maps

2) Surrogate Models

Surrogate models approximate complex models using simpler, interpretable models, such as decision trees or linear models.

Examples:

- Global surrogate decision trees
- Local linear models (LIME)

3) Example-Based Explanations

These explanations justify predictions using similar instances or prototypes from the dataset.

Examples:

- k-Nearest Neighbors explanations
- Prototypes and criticisms
- Case-based reasoning

4) Counterfactual Explanations

Counterfactuals describe minimal changes to input features that would alter the model's prediction.

Examples:

- "If income were increased by 10%, the loan would be approved."

5) Visualization-Based Explanations

Visualization techniques present explanations using graphs, heatmaps, or attention maps.

Examples:

- Attention heatmaps
- Activation maps (Grad-CAM)
- Feature interaction plots

Table 4.Summary of Explainable AI Taxonomy

XAI Category	Description	Representative Techniques
Feature-Attribution Methods	Assign importance scores to input features	SHAP, Integrated Gradients, Saliency Maps
Surrogate Models	Approximate complex models with interpretable ones	Decision Trees, LIME
Example-Based Explanations	Explain predictions using similar or representative instances	k-NN, Prototypes, Case-Based Reasoning
Counterfactual Explanations	Identify minimal changes to alter predictions	What-if analysis, Counterfactual rules
Visualization-Based Explanations	Use visual tools to interpret model behavior	Grad-CAM, Heatmaps, Interaction Plots

IV. REVIEW OF EXPLAINABLE AI TECHNIQUES

This section reviews widely used Explainable Artificial Intelligence (XAI) techniques, highlighting their underlying principles, advantages, and limitations. The methods are selected to cover model-agnostic, model-specific, local, global, and inherently interpretable approaches.

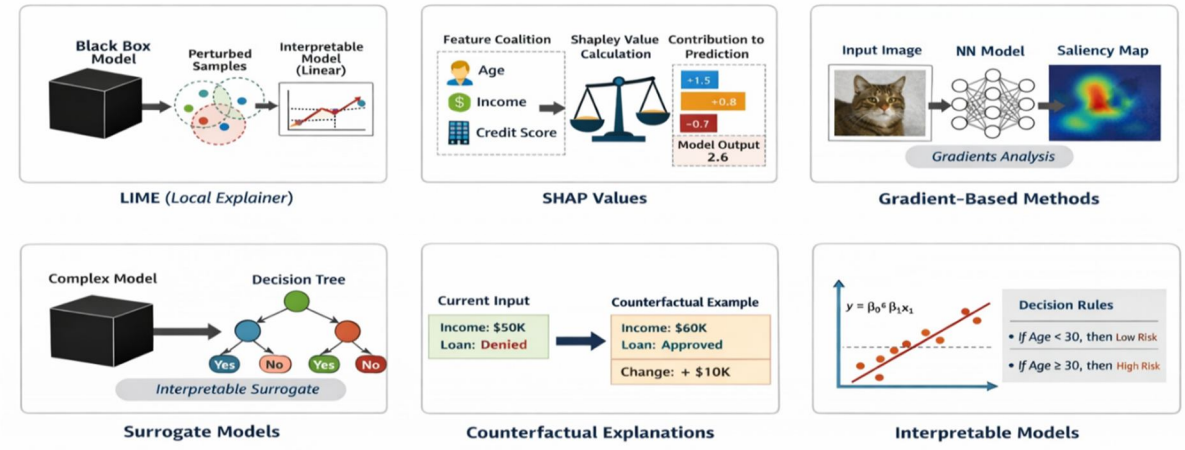


Fig 2. XAI Techniques

A. *Local Interpretable Model-Agnostic Explanations (LIME)*

LIME uses an interpretable surrogate model, usually a linear model, to approximate the behavior of a complex black-box model locally in order to explain individual predictions. The surrogate is taught to imitate the original model in that limited area by creating perturbations around the instance of interest.

Local Interpretable Model-Agnostic Explanations (LIME) approximates the behavior of a sophisticated black-box model in the local neighborhood of a particular instance in order to explain individual predictions. In order to observe the matching predictions from the original model, it creates perturbed samples around the instance of interest. These samples are then used to train an interpretable surrogate model, usually a linear model, to approximate the local decision boundary. The most important characteristics influencing the forecast are highlighted by the learnt surrogate model.

B. *SHapley Additive exPlanations (SHAP)*

SHapley Additive exPlanations (SHAP) is a feature-attribution method based on cooperative game theory. It assigns a Shapley value to each feature, indicating its contribution to the model's prediction, by considering all possible combinations of feature coalitions. Because it ensures attributes like fairness, consistency, and local accuracy, SHAP is a widely utilized and theoretically sound explanation technique for both global and local interpretability. In accordance with cooperative game theory, SHAP assigns a Shapley value to each attribute, signifying its contribution to the prediction. It ensures consistency and equity by considering all possible feature coalitions.

C. *Gradient-Based Methods*

Gradient-based approaches examine how modifications to input features impact the model's output in order to explain predictions. They rely on gradients of the output with regard to input features and are mostly used in deep neural networks. Techniques for gradient-based explanation look at how small changes to input features affect the model's output. By computing gradients of the result with respect to the input features, these techniques ascertain the sensitivity of the prediction. Gradient-based methods, which are mostly used to deep neural networks, are helpful for characterizing image, text, and audio models. Typical examples include saliency maps and integrated gradients.

D. *Surrogate Models*

Surrogate models use more straightforward, interpretable models, like decision trees or linear regressors, to approximate more complex models. To replicate the behavior of the original model, they can be trained locally or globally. Surrogate models use simpler, easier-to-understand models, like decision trees or linear regressors, to imitate complicated black-box models. These surrogate models can be trained locally to explain particular predictions or globally to reflect the general behavior of the original model. Surrogate models offer insights into decision-making while preserving interpretability by imitating the behavior of the black-box model.

E. *Counterfactual Explanations*

Counterfactual explanations provide useful information for decision-making by identifying small modifications to input features that would modify a model's forecast. Counterfactual explanations pinpoint the smallest adjustments needed to input characteristics in order to change a model's prediction. By addressing "what-if" concerns, such as how an outcome could be altered to obtain a desired result, they offer practical and intuitive insights. In regulatory compliance, fairness analysis, and decision-support systems, counterfactual explanations are very helpful.

F. *Inherently Interpretable Models*

Inherently interpretable models are designed to be transparent by nature, allowing explanations to be directly derived from their structure without post-hoc techniques. Rule-based systems, decision trees, and linear regression are a few examples of models that are highly interpretable and dependable, but sometimes they perform worse in terms of prediction than complex black-box models. Because inherently interpretable models are transparent by nature, explanations can be obtained immediately from their internal structure without the need for post-hoc methods.

Table 5: Taxonomy-Aligned Comparison of Explainable AI Techniques

Method	Explanation Mechanism	Core Idea	Key Strengths	Key Limitations
LIME	Surrogate model	Approximates a black-box model locally using an interpretable model	Simple, intuitive, applicable to any model	Unstable explanations, sensitive to sampling
SHAP	Feature-attribution	Uses Shapley values to assign feature contributions	Theoretically grounded, consistent explanations	Computationally expensive, feature dependence issues
Integrated Gradients	Feature-attribution	Computes attribution by integrating gradients from a baseline	High fidelity, suitable for deep networks	Requires differentiable models, baseline selection
Grad-CAM	Visualization-based	Uses gradients to generate class-specific activation maps	Intuitive visual explanations for images	Limited to CNNs, qualitative interpretation
Surrogate Models	Surrogate model	Mimics complex models using interpretable approximations	Human-readable, global insights	Approximation errors reduce fidelity
Counterfactual Explanations	Counterfactual	Identifies minimal input changes to alter predictions	Actionable, user-centric explanations	May produce unrealistic counterfactuals

V. EVALUATION METRICS AND QUALITY OF EXPLANATIONS

A key problem in Explainable Artificial Intelligence (XAI) is assessing the quality of explanations. The quality of an explanation is multifaceted and frequently subjective, in contrast to predicted performance. This section highlights both established metrics and unresolved issues while reviewing the main evaluation criteria used to evaluate the efficacy, dependability, and usability of XAI techniques.

A. Fidelity and Faithfulness

The degree to which an explanation accurately captures the actual behavior of the underlying model is known as fidelity, or faithfulness. If the model's output changes in tandem with the modifications the explanation suggests, the explanation is said to be loyal. Removing highly ascribed traits, for instance, should have a major effect on predictions.

Methods of Evaluation:

- Tests for perturbation or feature removal
- Consent between original models and surrogate models
- Sensitivity analysis

Challenges:

- Human interpretability is not guaranteed by high fidelity.
- Some explanations may appear plausible but be unfaithful

B. Stability and Robustness

Robustness quantifies resistance to noise or adversarial shifts, whereas stability relates to the consistency of explanations when inputs are mildly disrupted. For similar inputs or small perturbations that don't affect predictions, reliable explanations shouldn't change significantly.

Methods of Evaluation:

- Calculating the variance of explanations under input perturbations.
- Verification of consistency between comparable data instances.
- Robustness testing against noise.

Challenges:

- The balance between stability and sensitivity.
- Explanations that are too consistent could conceal significant model behavior.

C. Human Interpretability

Human interpretability evaluates how simple it is for people to comprehend, believe, and act upon explanations. The intended audience (such as domain experts versus end users) and the way explanations are presented affect interpretability.

Methods of Evaluation:

- Surveys and user studies
- Task-based assessment (response time, decision correctness)
- Analysis of cognitive burden and qualitative feedback.

Challenges:

- Subjective and context-dependent
- Standardization across areas is challenging.

D. Computational Efficiency

The time and resources needed to produce explanations are measured by computational efficiency. For large-scale or real-time applications, like online decision support systems, effective explanation techniques are crucial.

Methods of Evaluation:

- Analysis of memory and runtime usage.
- Testing scalability as data size or feature dimensions increase
- The overhead of explanation generation is compared.

Challenges:

- Efficiency and explanation quality trade-offs
- Exorbitant computing costs for theoretically based approaches (like SHAP)

E. Limitations of Current Evaluation Approaches

Despite significant progress, existing evaluation methods face several limitations:

- The absence of uniform metrics and standards •
- An excessive dependence on proxies rather than explanations based on the facts
- The challenge of striking a balance between usability, interpretability, and integrity
- Insufficient attention to domain-specific needs

These challenges highlight the need for hybrid evaluation frameworks that combine quantitative metrics, human-centered evaluation, and domain knowledge to assess explanation quality comprehensively.

VI. APPLICATIONS OF EXPLAINABLE ARTIFICIAL INTELLIGENCE

For real-world AI applications to be transparent, trustworthy, and accountable, explainable AI is essential. Opacity in decision-making can raise ethical, legal, and safety issues in high-stakes situations. The main application areas where XAI is crucial for the responsible use of AI are highlighted in this section.

A. Healthcare and Medical Diagnosis

Explainable Artificial Intelligence plays a major role in modern healthcare systems by helping medical professionals understand the reasoning behind AI-generated decisions. AI models are widely used for disease prediction, medical image analysis, patient monitoring, and treatment recommendation. XAI provides clear explanations for these predictions, allowing doctors to verify whether the conclusions are medically reliable. For example, in cancer detection systems, explainable models can highlight the affected areas in medical scans and justify why a disease is predicted. This improves trust, increases patient safety, and supports better clinical decision-making.

B. Finance and Banking

In the financial sector, AI systems are used for credit scoring, loan approval, fraud detection, and risk assessment. Since financial decisions directly affect customers, transparency becomes essential. XAI helps banks and financial institutions explain why a loan was approved or rejected and how risk levels were calculated. This reduces the chances of unfair or biased decisions and ensures compliance with financial regulations. By providing understandable explanations, XAI increases customer confidence and supports ethical financial practices.

C. Cybersecurity and Threat Detection

Cybersecurity systems use AI to identify malware, network intrusions, phishing attacks, and suspicious activities. However, security analysts need to understand why a particular activity is classified as a threat. XAI provides detailed explanations about detected anomalies and attack patterns, enabling experts to respond more effectively. It also helps reduce false alarms and improves the efficiency of security operations. As cyber threats continue to grow, explainable AI becomes important for building trustworthy and intelligent security systems.

D. Education and Intelligent Learning Systems

Educational institutions use AI for personalized learning, automated grading, and student performance analysis. XAI helps students and teachers understand how learning recommendations and performance evaluations are generated. For example, an intelligent tutoring system can explain why certain learning materials are recommended to a student based on their progress and weaknesses. This transparency improves the learning experience and helps educators make informed academic decisions.

E. Human Resource Management

Organizations use AI systems for recruitment, resume screening, employee monitoring, and performance evaluation. XAI ensures that hiring decisions are fair and understandable by explaining why a candidate was selected or rejected. It helps organizations detect bias in recruitment systems and supports equal employment opportunities. Transparent AI systems also improve employee trust and organizational accountability.

F. Defense and National Security

Defense organizations use AI for surveillance, threat analysis, battlefield monitoring, and intelligence gathering. Since military decisions are highly sensitive, human operators must clearly understand the reasoning behind AI-generated outputs. XAI ensures transparency in mission-critical systems and supports better strategic planning. It also reduces the risks associated with fully automated decision-making in defense operations.

VII. CHALLENGES AND LIMITATIONS OF EXPLAINABLE ARTIFICIAL INTELLIGENCE

Despite significant progress, Explainable Artificial Intelligence faces several fundamental challenges that limit its effectiveness, reliability, and adoption in real-world systems. This section critically examines key limitations and open issues in current XAI research.

A. Accuracy vs. Interpretability Trade-off

The trade-off between interpretability and forecast accuracy is one of the most enduring issues in XAI. While intrinsically interpretable models may compromise performance, highly accurate models—like deep neural networks and ensemble approaches—are frequently intricate and challenging to comprehend.

In high-stakes applications where accuracy and openness are crucial, this trade-off makes choosing a model more difficult.

B. Lack of Standard Definitions and Frameworks

In AI, explainability and interpretability have no agreed-upon meaning. These terminologies are used inconsistently among studies, which leads in inconsistent evaluation standards and uncomparable outcomes. The systematic evaluation and comparison of XAI techniques is hampered by the lack of defined taxonomies, benchmarks, and evaluation procedures.

C. *Subjectivity in Explanations*

Interpretability is subjective by nature and is contingent upon the end-user's expectations, experience, and background. A layperson or domain specialist might not understand an explanation that makes sense to a data scientist.

It is challenging to create explanations that are universally effective and to objectively assess the quality of explanations because of this subjectivity.

D. *Adversarial Manipulation of Explanations*

It has been demonstrated in recent research that explanations themselves are manipulable. Models may be purposefully created to conceal biased or dangerous behavior by producing false but believable explanations. Such hostile manipulation casts doubt on the validity of post-hoc explanation techniques and erodes confidence in XAI.

E. *Scalability and Real-Time Constraints*

A large computational expense is associated with many explanation methods, especially those with strong theoretical assurances. This restricts their use in real-time or large-scale systems like autonomous systems and online recommendation systems. Research on striking a balance between computational efficiency and explanation quality is still ongoing.

F. *Ethical and Legal Challenges*

Fairness, accountability, and openness are among the ethical and legal factors that are intimately linked to XAI. Explanations, however, may unintentionally reveal private information, strengthen prejudices, or provide a false impression of justice. Additionally, there is still uncertainty surrounding the legal criteria for explainability, which affects businesses using AI systems in regulated fields.

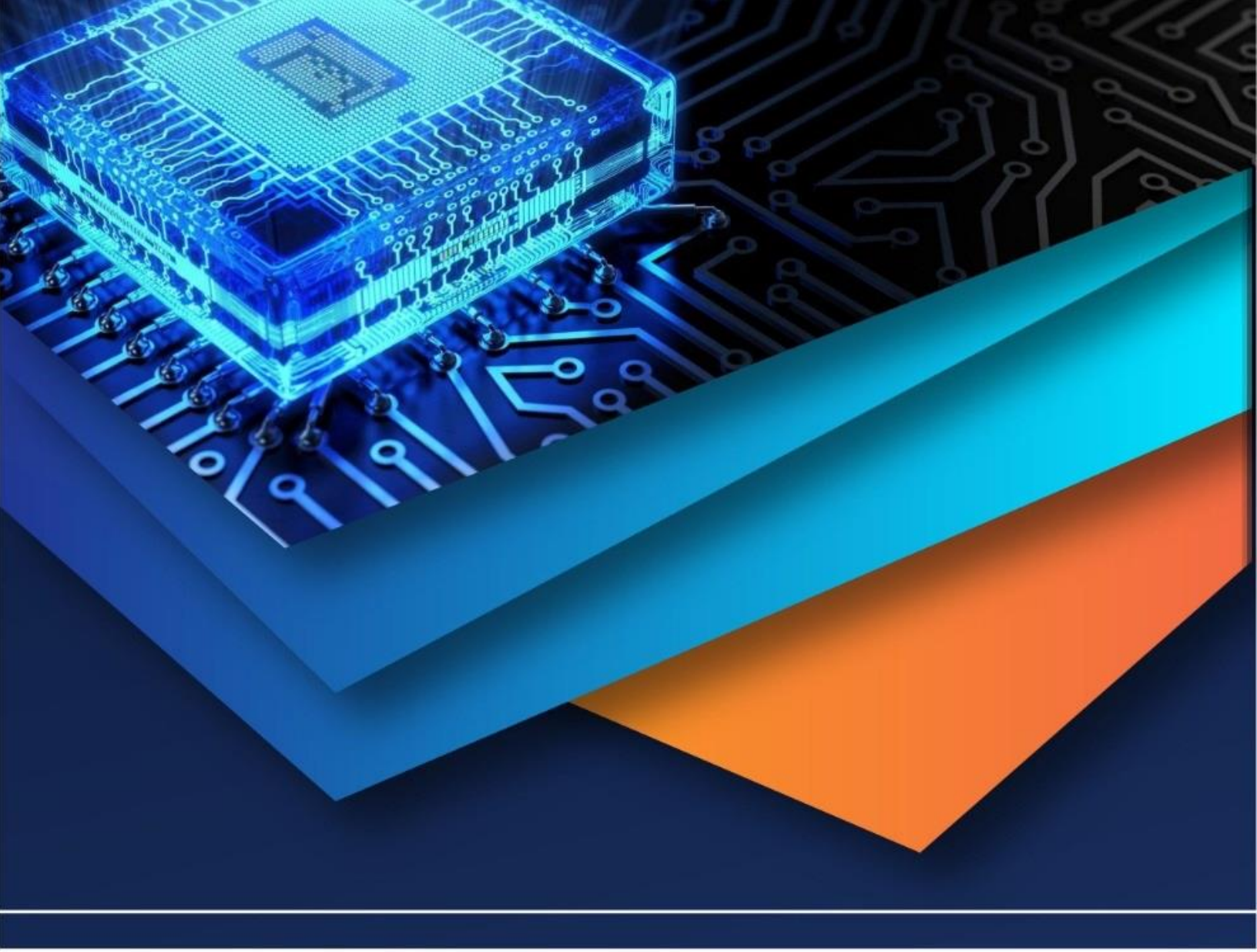
VIII. CONCLUSION

This study presented a comprehensive review of Explainable Artificial Intelligence (XAI), including its concepts, methods, evaluation techniques, applications, challenges, and future research directions. A structured taxonomy was introduced to categorize XAI approaches based on explanation type, model dependency, and scope of interpretation. The review showed that no single XAI method is universally suitable, as effectiveness depends on user needs and application domains. The paper also emphasized the importance of balancing interpretability, accuracy, stability, and usability. Applications in healthcare, finance, law, and autonomous systems demonstrated the significance of XAI in building trustworthy AI. Future research should focus on human-centered and real-time explainability solutions.

REFERENCES

[1] L. Longo, M. Brcic, F. Cabitza, et al., “Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions,” arXiv preprint arXiv:2310.19775, 2023.
[2] M. Saarela and V. Podgorelec, “Recent Applications of Explainable AI (XAI): A Systematic Literature Review,” Applied Sciences, vol. 14, no. 19, p. 8884, 2024.
[3] A. Ghasemi, S. Hashtarkhani, D. L. Schwartz, and A. Shaban-Nejad, “Explainable Artificial Intelligence in Breast Cancer Detection and Risk Prediction: A Systematic Scoping Review,” arXiv preprint arXiv:2407.12058, 2024.
[4] Z. M. Altukhi, S. Pradhan, and N. Aljohani, “A Systematic Literature Review of the Latest Advancements in XAI,” Technologies, vol. 13, no. 3, p. 93, 2025.
[5] D. E. Mathew, D. U. Ebem, A. C. Ikegwu, and P. E. Ukeoma, “Recent Emerging Techniques in Explainable Artificial Intelligence to Enhance the Interpretable and Understanding of AI Models for Human,” Neural Processing Letters, vol. 57, no. 16, pp. 1–25, 2025.
[6] G. Paliwal, A. Kumar, K. P. Sharma, and D. Bhargava, “Transformative Impact of Explainable Artificial Intelligence: Bridging Complexity and Trust,” Discover Artificial Intelligence, vol. 5, no. 51, pp. 1–18, 2025.
[7] A. Choudhari, T. Ambhore, and N. Mehendale, “Explainable AI: Bridging the Gap Between Machine Intelligence and Human Understanding,” SSRN Electronic Journal, 2025.
[8] J. Sanderson, H. Mao, and W. L. Woo, “GradCFA: A Hybrid Gradient-Based Counterfactual and Feature Attribution Explanation Algorithm for Local Interpretation of Neural Networks,” IEEE Transactions on Artificial Intelligence, vol. 6, no. 10, pp. 2575–2587, 2025.
[9] V. Swamy, D. Romano, B. Srinivasa Desikan, O.-M. Camburu, and T. Käser, “iLLuMinaTE: An LLM-XAI Framework Leveraging Social Science Explanation Theories Towards Actionable Student Performance Feedback,” in Proc. AAAI Conf. Artificial Intelligence, vol. 39, no. 27, pp. 28431–28439, 2025.
[10] “Explainable Artificial Intelligence Models in Intrusion Detection Systems,” Engineering Applications of Artificial Intelligence, vol. 144, p. 110145, 2025.
[11] “A Review of Explainable Artificial Intelligence from the Perspectives of Challenges and Opportunities,” Algorithms, vol. 18, no. 9, p. 556, 2025.
[12] “Next-Gen Explainable AI (XAI) for Federated and Distributed Internet of Things Systems: A State-of-the-Art Survey,” Future Internet, vol. 18, no. 2, p. 83, 2026.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2016, pp. 1135–1144.
[14] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 4765–4774.
[15] R. R. Selvaraju, M. Cogswell, A. Das, et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017, pp. 618–626.
[16] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," in Proc. Int. Conf. Machine Learning (ICML), 2017, pp. 3319–3328.
[17] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," Information Fusion, vol. 58, pp. 82–115, 2020.
[18] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," Entropy, vol. 23, no. 1, p. 18, 2021.
[19] C. Molnar, Interpretable Machine Learning. Lulu.com, 2020.
[20] T. Miller, "Explanation in Artificial Intelligence: Insights from the Social Sciences," Artificial Intelligence, vol. 267, pp. 1–38, 2019.
[21] C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206–215, 2019.
[22] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR," Harvard Journal of Law & Technology, vol. 31, no. 2, pp. 841–887, 2018.
[23] D. Guidotti, A. Monreale, S. Ruggieri, et al., "A Survey of Methods for Explaining Black Box Models," ACM Computing Surveys, vol. 51, no. 5, pp. 1–42, 2018.
[24] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, 2018.
[25] D. Gunning and D. Aha, "DARPA's Explainable Artificial Intelligence (XAI) Program," AI Magazine, vol. 40, no. 2, pp. 44–58, 2019.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY