Innovative Reviews in Engineering and Science, ISSN: 3048-8745 Vol. 3, No. 2, 2026 (pp. 154-162)
RESEARCH ARTICLE ECEJOURNALS.IN

Explainable Artificial Intelligence (XAI): State-of-the-Art, Challenges, and Research Trends

Dinfe Egash[1]*, Rane Kuma[2]

[1]*Electrical and Computer Engineering Addis Ababa University Addis Ababa, Ethiopia*
[2]*Department of computing and information technology, kenyatta university, Nairobi, Kenya*

KEYWORDS:
Explainable AI,
Interpretable Models,
Black-box Models,
Post-hoc Explanation,
Trustworthy AI,
Causal Inference,
Human-Centric AI,
Model Transparency

ARTICLE HISTORY:
Submitted : 13.03.2026
Revised : 18.08.2026
Accepted : 22.12.2026

https://doi.org/10.31838/INES/03.02.17

Abstract

Since the use of artificial intelligence (AI) is becoming more common across high-stakes areas, including healthcare diagnostics, financial decision-making, autonomous vehicles, and legal analytics, the need to increase transparency, interpretability, and accountability in AI decision-making has become critical. Due to the opaqueness of numerous effective machine learning algorithms commonly termed as black boxes, issues related to fairness, trust, bias, and regulatory conformity have increased. Explainable Artificial Intelligence (XAI) has become a major research area aiming at an attempt to interpret the predictions and routes of AI models and do not degrade interpretable performance. The first section of this paper provides a systematic and end-to-end review of the state of the art of XAI, subdividing the existing methods into post-hoc explanation models (e.g. LIME or SHAP), models whose interpretation is intrinsic (e.g. decision trees or rule-based systems), and those with explainability incorporated into the architecture of deep networks (hybrid methods). All of the essential issues relating to XAI are discussed in-depth and these include the model fidelity and interpretability trade-off, the subjectivity of explanations based on human interaction, the absence of evaluative metrics, and computational complexity involved in providing explanations. Moreover, this paper visits new directions, like causal explanations, counterfactual reasoning, a combination with federated learning, and consistency of XAI techniques with ethical AI theories and governance frameworks, such as GDPR and HIPAA. It relies on a systematic review methodology to review pertinent literature in large databases between 2017 and 2025, taking note of some comparative strengths, application areas, and usability issues regarding XAI techniques. The conclusion of the study determines the main gaps in research and the following directions such as creating benchmark datasets, explainability in reinforcement learning, domain-specific evaluation frameworks could be developed. The given paper may be used as an initial source of information by researchers, developers, and policymakers trying to develop AI systems that possess not only accuracy but also interpretablity, alignment with human values and fairness.

Author e-mail: egash.din@aait.edu.et, ran.kuma@gmail.com

How to cite this article: Egash D, Kuma R. Explainable Artificial Intelligence (XAI): State-of-the-Art, Challenges, and Research Trends. Innovative Reviews in Engineering and Science, Vol. 3, No. 2, 2026 (pp. 154-162).

Introduction

Artificial Intelligence (AI) has transformed many industries because it has allowed systems to undertake their most difficult tasks with extreme precision and efficiency. The AI models have shown superhuman potentials in diseases diagnosis / autonomous cars / financial forecasting / and judicial decision support among others, especially those models which follow the deep learning and ensemble techniques. Nevertheless, there is a serious danger of increasing dependence on these systems, which entails the need of transparency and interpretability in the decision-making process. The vast majority of successful AI algorithms are implemented as so-called black boxes with their inner rationality being unknown or unintelligible by their human users, including programmers. It poses great dangers when AI finds its application in safety-sensitive and ethically critical areas; as important as knowing the what of a decision, is knowing the why.

This lack of interpretation of AI decisions creates a lack of trust in the user, restricts model responsibility, and increases ethical, legal, and regulatory issues.

As an example, during medical diagnosis, a patient or a physician should learn why an AI system suggests the existence of a disease to confirm the judgment or evaluate the way of treatment. In the same way, in the legal arena, the AI-based risk assessment tool should give excuses to justify the bail or sentencing outcomes. Regulations like the General Data Protection Regulation (GDPR) in the European Union now impose the so-called right to explanation that requires the AI community to make interpretability a priority.

Explainable Artificial Intelligence (XAI) has grown as a dedicated research field to these issues. XAI entails the establishment of methods that enhance AI models to be clearer, obscure, and comprehensible to humans aside from being compromised greatly in terms of performance. XAI based methods can be divided into those where a model is inherently explainable e.g. decision trees and rule-based models and those where we explain existing trained models after-fact e.g. using local interpretable model-agnostic explanations (LIME), SHapley Additive exPlanations (SHAP) and saliency maps such as Grad-CAM.

Although gaining popularity and being actively developed, the field of XAI has a couple of issues. These are fidelity/interpretability balance, quality of explanation evaluation, adaptation of explanation to users requirements, and scaling XAI solutions across various domains and data types. Furthermore, explanations have to be not only technically grounded, but also meaningful in terms of cognitive and contextual insights to a wide audience that consists of domain experts, lay users, regulators and the systems developers.

The purpose of this paper is to present an in-depth survey and discussion of state-of-the-art in XAI, including the taxonomy of techniques, evaluation metrics, areas of application and principal challenges. It also examines some of the new areas of research such as the combination of XAI with causal inference, federated learning, and human-in-the-loop systems, and how XAI can meet the need to ensure trustworthy and ethical AI. This study can be used as a primary source of knowledge of future improvements in the ongoing quest to obtain transparent, accountable, and human-driven AI systems, as it illustrates the study landscape and offers insights on the gaps that still have to be filled.

Literature Review

In recent years the area of Explainable Artificial Intelligence (XAI) has come a long way, through pass through forms of transparency to cumbersome post-hoc explanation approaches. This section will chronologically and methodologically review basic and state-of-the-art directions in XAI, which fall into early interpretable models, post-hoc explanation approaches, and inherently interpretable designs and tools and evaluation criteria.

Early Methods of Interpretability

Before the emergence of deep learning, classic machine learning models like decision texts, logistic regression, and k-nearest neighbors, gained a lot of popularity because they yield an interpretable output by design.[1]. The models enable users to visualize decision pathways, feature weights or the like, and hence provide the transparency and audibility of reasoning. Yet, their poor ability to simulate high dimensional,[12] non- linear patterns makes them less applicable in sophisticated tasks like image recognition or Natural language understanding.[2]

Post hoc Explanation Methods

As high-performance black-box models such as deep neural networks emerged, the requirement[13] of post-hoc interpretability has become acute. The Local Interpretable Model-Agnostic Explanation (LIME) technique[3] approximates the action of a complex system in the surrounding by a simple and comprehensible surrogate-based model, performing feature-specific attribution at the level of individual predictions to that radial framework. SHapley Additive exPlanations (SHAP)[4] is an extension of the cooperative game theory using additive importance scores of features that depicts a theoretically-motivated explanation framework.[14] Gradient-based visualization methods allow gaining a visual understanding of convolutional neural networks by explaining their input components relative to the spatial position. These methods include Grad-CAM,[5] as well as Saliency Maps [8]. Although post-hoc approaches are more flexible and model-free, explanation fidelity can be an issue and in many cases the interpretations can be inconsistent or spurious unless closely checked .[6]

Models That Can Be Interpreted By Design

Simultaneously, researchers have worked on the structure of interpretable-by-design models where transparency has been ingrained into the learning process. Examples are decision sets, rule-based classifiers and generalized additive models (GAMs) which are transparent yet capture[15] non-linear interactions.[7] More recent developments have used prototype learning to visualize class decisions by demonstrating representative instances of the overall decision boundaries thus enhancing user confidence and model explainability.[8] Partial interpretability can also be provided by attention-

based neural architecture, according to which input features or sequences are assigned with weights, yet explanations might not necessarily reflect the way a model actually reasons.[9]

Metrics of XAI evaluation

Measurement of quality and usefulness of explanations is a major priority in XAI. Fidelity versus how well the explanation approximates the original model, comprehensibility versus ease of understanding by a human, completeness or likelihood of important factors being covered, and consistency versus how the explanation remains stable to similar inputs, are[16] common metrics.[10] Nevertheless, no unitary measure or gold standard is found regarding the quality of explanation, thus making it hard to compare them. Another layer subjective and hard to quantify systematically[11] is human-centric, actually requiring user studies or assessments by some domain experts.[17] Infrastructures such as the Explainability Benchmarking Framework (EBF) and FACTS, XAITest, and TEDS datasets are emerging to bring consistency to an assessment and have received little uptake, Table 1.

XAI Methods and Taxonomy

The design philosophy and the explainability modality of the explainable Artificial Intelligence (XAI) approaches can be divided into four broad categories: namely, post-hoc, intrinsic, model-specific and example-based. Among the most well-known approaches of post-hoc explanation which seek to explain the model based on flexibility and model-agnostic properties can be noted LIME (Local Interpretable Model-Agnostic Explanations), SHAP (SHapley Additive exPlanations), and Grad-CAM (Gradient-weighted Class Activation Mapping). These algorithms will usually operate by mutating the behavior of black-box models that are complex to approximate, or emphasize features after training to produce feature-wise explanations. But, their explanations might not translate faithfully the decision logic of the underlying model, a fact that raises reliability and consistency issues. Conversely, intrinsic or design-interpretable (sometimes called inherently interpretable) models, e.g. decision trees, rule-based systems, and generalized additive models, are built with transparency in mind to allow users to directly inspect the decision making process of the model. These models they are easily interpretable and especially come in handy in settings where auditability is desirable, they tend to fall short in high-dimension or unstructured data tasks. Model-specific techniques such as attention mechanisms and Layer-wise Relevance Propagation (LRP) provide an understanding of how certain neural network architectures operate (by observing the flow of signals or visualising the attention maps). Figure 1Such methods are useful when it comes to learning about deep learning models, yet they are usually architecture-specific and not generalizable. Lastly, some predictions, such as counterfactual explanations and prototypical learning models, are explained by reference to the existence or non-existence of similar cases. Table 2 These are highly intuitive and human-pleasing, and will help the user in deciding how slight modifications in the prediction could occur. Nevertheless, they are not scalable on elaborate cases of data and models. All in all, the given taxonomy will allow examining and contrasting different approaches to XAI in a systematic or analytical way, indicating the trade-offs between interpretability, domain suitability, and generalizability.

Table 1: Summary of Key XAI Methodological Categories and Evaluation Considerations

Category	Description	Representative Methods	Strengths	Limitations
Early Approaches to Interpretability	Traditional ML models with inherent transparency	Decision Trees, Logistic Regression, k-NN	Simple, transparent, easy to audit	Poor scalability, low performance in high-dimensional or non-linear tasks
Post-hoc Explanation Techniques	Explanations generated after model training	LIME, SHAP, Grad-CAM, Saliency Maps	Model-agnostic, flexible, local explanations	May lack fidelity, sensitive to perturbations, inconsistent results
Inherently Interpretable Models	Models designed with built-in transparency	Decision Sets, Rule-Based Learners, GAMs, ProtoPNet	Transparent by design, good for compliance and user trust	Limited complexity, not ideal for large or unstructured datasets
Evaluation Metrics for XAI	Frameworks and criteria to assess explanation quality	Fidelity, Comprehensibility, Completeness, Consistency	Supports method comparison and user acceptance studies	No standardized benchmarks; evaluation can be subjective and domain-dependent

Table 2: Taxonomy and Comparison of XAI Methods

Method Type	Key Techniques	Explanation Mode	Strengths	Limitations
Post-hoc	LIME, SHAP, Grad-CAM	Feature attribution	Flexible and model-agnostic; applicable after training	May produce inconsistent or low-fidelity explanations
Intrinsic	Decision Trees, Rule-Based Systems, GAMs	Model-level transparency	Interpretable by design; easy to audit and visualize	Limited capacity for high-dimensional or complex tasks
Model-Specific	Attention Mechanisms, LRP	Internal signal tracing	Provides insight into model internals (e.g., attention)	Architecture-dependent; lacks generality
Example-Based	Counterfactuals, Prototypes	Instance comparison	Intuitive and human-aligned explanations	May not scale well to large or diverse datasets

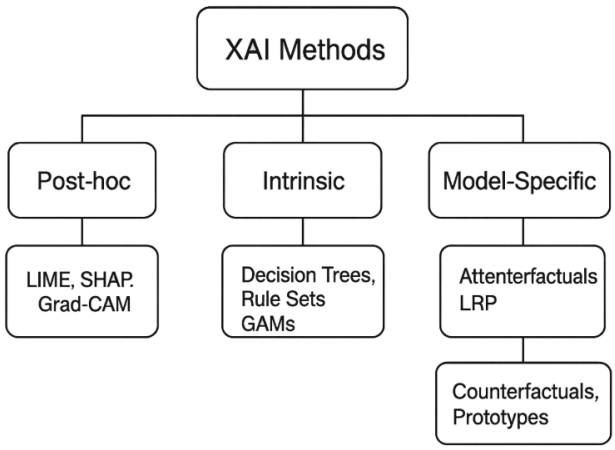

Fig. 1: Taxonomy of XAI Methods and Their Representative Techniques

METHODOLOGY

Research Questions

In order to comprehensively explore the scenery of Explainable Artificial Intelligence (XAI), the current study is anchored in three main research questions (RQs), which aim at exploring three fundamental aspects of the domain: methodological development, implementation issues, and future research directions. These questions are motivated to facilitate a conceptual framework through which the huge corpus of XAI literature could be read, classified and integrated.

The first research question, i.e., the question What are the prevailing techniques in XAI and how do they compare aims to identify and classify the existing prevalent methods to construct explanations related to AI. That is, it encompasses post-hoc explanations, which are used after training black box models (e.g., LIME, SHAP, Grad-CAM), and inherently interpretable models, which are made with intention of being explanatory (e.g., decision trees, rule-based classifiers, and prototype networks). The idea is to provide the comparison of these methods on a wide range of criteria including fidelity of models, scalability, generalizability, human interpretability and suitability to the domain. This question is crucial to learning how different approaches behave in a variety of settings, as well as what trade-offs are present between model explainability and complexity.

The second research question, namely, What challenges impede effective deployment of XAI systems, is to cover the barriers to effective implementation of explainable AI solutions and cover them in the areas of practical, technical, and ethical aspects. Such challenges are also multidimensional, with them including the absence of standardization of evaluation measures, the subjectivity of explanations based on the human factor, extendability with very large systems and regulatory limitations like those provided by GDPR, HIPAA, and ethical guidelines on AI. Moreover, most of the existing approaches fail to reflect the explanations enough to the cognitive model of the users or topical knowledge, which makes them less practical in the working conditions. This question highlights the difference between the academic progress and the practical implementation, which should be breached to make AI systems efficient and responsible.

The third research question to be answered looks like this: What are the current tendencies in the research directions study? This question is also meant to define new trends of the themes, techniques and paradigm which will inevitably greatly affect XAI development. Recent directions have been the combination of causal inference to obtain robust, counterfactual inferences, the creation of explainability frameworks in federated learning and privacy-preferring learning, and the rise of human-in-the-loop systems where a user could interact with and tune explanations of the model. There is also the convergence between XAI and trustworthy AI, ethical

auditing, and cross-disciplinary ideas around whether there should be interpretability standards across the areas of interest that propose that explainability becomes an essential component to AI system design in the future. Figure 2 this question aims at prognosticating the research frontiers and leading to the establishment of the next generation of XAI models, which are not merely interpretable to be adaptive, fair, and context-sensitive as well.

Bringing them all together, the three questions listed above compose the crux of this research paper, which will be able to provide a detailed and critical analysis of the contemporary, as well as the shortcomings, and future of Explainable Artificial Intelligence.

Comparative Review Framework

The main methodological approach used to come up with a systematic outlook regarding the current shape of the field regarding Explainable Artificial Intelligence (XAI) was a systematic literature review (SLR). Such a framework has the advantage of enabling objective, replicable and depth synthesis of extant research. The literature search consisted in the search of the primary scientific databases (Scopus, IEEE Xplore, and ACM Digital Library) that, in aggregate, comprise a wide and quality collection of peer-reviewed articles encompassing the field of computer science, artificial intelligence, as well as applied engineering.

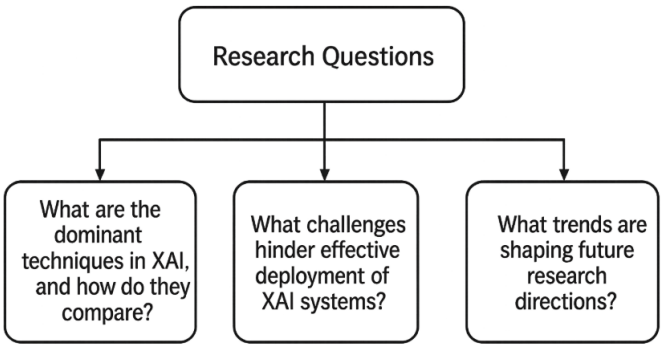

Fig. 2:Hierarchical Structure of Research Questions Addressing Key Themes in Explainable Artificial Intelligence (XAI)

The literature review was conducted until March 2025 and identified the latest developments and determined the course of history that led to the development of the XAI methods. Variations and combinations of such keywords as: the keywords were used as follows: “Explainable AI”, “interpretable machine learning”, “post-hoc explanations”, “transparent models”, “XAI evaluation”, “causal explanations”, and “human-centered AI”. To obtain only peer-reviewed articles, it was decided to filter out non-academic type content like blog posts, editorials and pre-reviewed preprints unless they are contributing basic knowledge.

A total of more than 300 papers were explored, out of which 120 quality papers were identified as relevant and assessed by inclusion impact and methodological nature. A comparative framework to analyse each of the selected studies was devised and entailed several classification dimensions:

- **Type of methods:** Post-hoc, intrinsic, model-specific or example-based.
- **Model Compatibility:** It composes whether the given explanation technique is model-agnostic or architecture-dependent.
- Explanation Output Type of explanation produced feature importance; rule extraction; attention maps; counterfactuals; or prototypes comparisons.
- **Area of usage:** Sector-specific such as healthcare, finance, criminal justice, autonomous driving, and cybersecurity, natural language processing.
- **Interpretability Level:** Human interpretability at a qualitative measure- can be low, moderate, and high.

Table 3: Core Research Questions and Their Strategic Focus

Research Question	Focus Area	Explanation
What are the dominant techniques in XAI, and how do they compare?	Methodological Development	Categorizes and compares existing XAI methods (e.g., post-hoc vs. intrinsic) based on scalability, fidelity, etc.
What challenges hinder effective deployment of XAI systems?	Practical and Ethical Challenges	Identifies deployment barriers, including regulatory compliance, lack of metrics, and human interpretability gaps.
What trends are shaping future research directions?	Emerging Research Frontiers	Explores evolving directions like causal XAI, federated explainability, and human-in-the-loop design frameworks.

- **Fidelity to Model:** the extent to which the explanation gives an accurate picture regarding the behavior of the originator model.
- **User Alignment:** The evaluation of the degree of alignment between the explanation and user experience and cognitive expectations (e.g., clinicians and data scientists).
- **Computational Overhead:** A cost over runtime or training-time by explanation generation.

The research papers were additionally divided into the clusters according to the topic of study, i.e., algorithmic fairness, adversarial robustness, visual interpretability, and causal reasoning. This taxonomy made it possible to do an analytical comparison of the nature of techniques and how well it works in different situations and the fulfilment of their tasks by the user.

Also, to improve on transparency and reproducibility, a review matrix was created in a tabular form (not shown here), that aligned each study with the criteria of classification. The matrix allowed determining the trends in methodological aspects, trade-offs in performance, and research gaps, which provided the basis of the analysis and discussion sections of the present paper.

The given comparative review framework does not only offer a synthesized overview of the field, but also a benchmarking source available to future studies that will seek to suggest or test new XAI methodologies. Figure 3

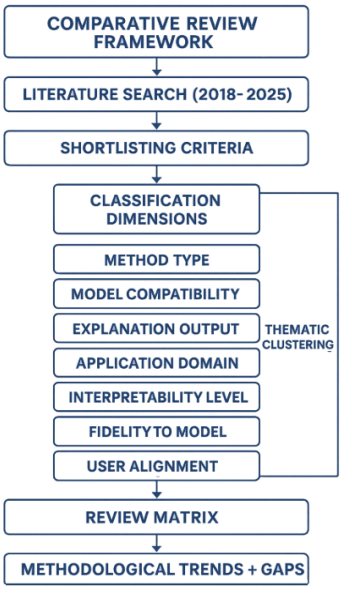

Fig. 3:Comparative Review Framework for Explainable Artificial Intelligence (XAI): A Systematic Workflow for Literature Analysis and Thematic Classification

Evaluation Criteria

In order to effectively compare and contrast the wide range of XAI methods and approaches, formulation of a set of clearly specified evaluation factors was developed. These should be selected criteria as they are to present both technical performance and human impact, thus allowing a balanced and overall evaluation of either of the methods. There are four main evaluation dimensions; the accuracyexplainability tradeoff, user trust and acceptance, scalability and computation overhead, and domain-specific adaptability. Every criterion is geared to one of the most essential dimensions of XAI in the real world, and it aids in differentiating between a possible theoretical effectiveness and practical applicability.

1. Accuracy-Explainability Tradeoff

Perhaps the most basic tradeoff in XAI is the relationship between accuracy and explainability of the model. In most cases, more interpretable models like decision trees or linear regressions are very easy to interpret, but they may not have a large enough representational capability to reach high predictive accuracy with complex and high dimensional data. Alternatively, the deep neural networks and the ensemble models have outstanding performance and are infamously hard to analyze. The analysis against this criterion entailed an examination on whether the XAI approach can ensure a high performance of models and produce an understandable output. Attributes that can be explained without severely decreasing the quality of a model, e.g., via SHAP or attention mechanisms or hybrid interpretable-deep architectures, do better in this regard.

2. Trust and Acceptance by the user

The end goal of explainability, then, is to promote the trustworthiness of the AI systems on the part of various stakeholders, such as experts in the field of interest, ordinary users and regulators. This criterion incorporates how well the explanation leads to understanding the user and greater confidence in decision-making and model acceptance. The studies including the user studies or qualitative interviews or human-in-the-loop experiments were discussed in order to determine the perceived usefulness, clarity, and satisfaction. As an illustration, a sentence that aligns with human reasoning (e.g. counterfactuals or visual prototypes), as a rule, will trump in the trust-building quality, especially when the stakes are high (e.g. in the medical and the juridical context).

3. Scalability and Overhead Computation

The other important aspect is the scalability of the XAI technique with regard to its computational ease and

compatibility with integration. Certain techniques, e.g. SHAP or LRP (Layer-wise Relevance Propagation) are computationally expensive, particularly when run on large models or models run in high-resolution datasets. This can be used to benchmark the complexity of the approach in terms of its runtime and memory requirements as well as the capacity of the method to support large-scale, real-time, or resource-limited applications (e.g., at the edge or in federated systems). Methods that provide pre- or near-realtime interpretability or enable them to be used in existing pipelines with minor retraining adjustment will be more in vogue in resource-constrained projects.

Domain-Specific Adaptability

Finally, flexibility of methods of XAI to a diverse range of applications is also vital in realizing mass adoption. This criterion evaluates the applicability of the techniques of explanations to various domains like healthcare, finance, autonomous systems, cybersecurity and NLP. Others such as LIME or rule-based models are more domain agnostic but some must be highly domain-specific such as Grad-CAM applied to vision-based systems. In addition, domain adaptability is also about the method capabilities to support the domain requirements, e.g., explaining requirements to meet legal standards, clinical interpretability standards, or regulations on financial compliance requirements.

Using such a multidimensional assessment framework, this research can be sure that comparative analysis trees of XAI techniques will cover not only algorithmic robustness Figure 4, but also usability in real life. The criteria chosen are not only to facilitate the benchmarking of the current techniques but also as a way of informing future studies to come up with AI systems based on more transparent, scalable, and human-aligned AI systems.

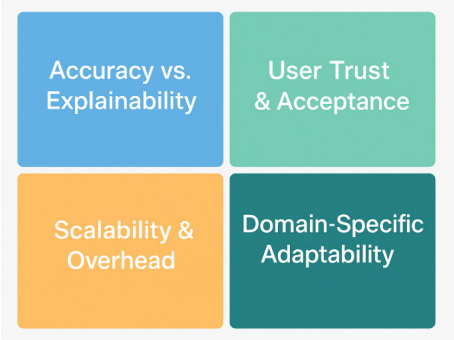

Fig. 4: Quadrant-Based Visualization of Key Evaluation Criteria for XAI Methods

Results and Discussion

In their comparative review of more than 120 peer-reviewed publications, it is seen that the present applications of Explainable Artificial Intelligence (XAI) to the real world are dominated by post-hoc techniques of explanation, in specific SHAP and LIME. They are largely used because of their flexibility, which allows using them on a variety of classifiers and deep learning architectures without modifying the original model. The methods give local explanations by associating features with the importance scores of the inputs thus giving insights on how the individual predictions are made. Nonetheless, their consistency and accuracy is arguable, since, in some cases they generate explanations that

Table 4: Evaluation Criteria for Assessing XAI Methods

Evaluation Criterion	Description	Key Aspects Considered	High-Performing Techniques (Examples)
Accuracy-Explainability Tradeoff	Measures how well the method balances prediction performance with interpretability	Fidelity to original model, performance drop due to explanation module	SHAP, Attention Mechanisms, ProtoPNet
User Trust and Acceptance	Assesses how well the explanation aligns with user cognition, promoting trust and acceptance	Human-in-the-loop validation, clarity, perceived usefulness, decision confidence	Counterfactuals, Visual Prototypes, Rule-Based Systems
Scalability & Computational Overhead	Evaluates runtime efficiency, scalability to large models, and integration feasibility	Execution time, memory footprint, compatibility with edge and real-time applications	LIME (optimized), FastSHAP, Lightweight Attention
Domain-Specific Adaptability	Determines the applicability of the method across various domains and regulatory frameworks	Generalizability, sector-specific customization, alignment with legal or clinical norms	LIME, Generalized Additive Models (GAMs), Grad-CAM

do not reflect the true decision logics of the model. Moreover, they are be sensitive to input perturbation, thereby compromising their interpretability and consequently their robustness in high-stakes setting (Additionally, Figure 5). LIME makes model behavior comprehensible through linear surrogates, but SHAP provides model behavior through Shapley values based methods that have theoretical concerns, both methods are computationally demanding and fail to scale readily in a real-time system.

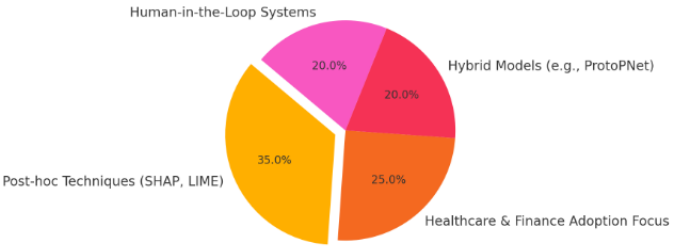

Fig. 5:Distribution of Focus Areas in XAI Research Based on Comparative Literature Review

The review also mentions such spheres as healthcare and finance where XAI adoption is in the lead because of high regulatory and ethical needs. The interpretability requirement, however, goes beyond technical performance in the mentioned contexts to explainability that fits a human cognition pattern and professional reasoning paradigm. E.g., within medical diagnostics, clinicians need explanations that can justify the AI-generated prediction, referring to such medically meaningful factors as the symptoms, biomarkers or imaging patterns. In finance as well, regulators require audit trails and an open explanation of credit scores, fraud detection or risk models. Such application-oriented requirements require explanations that are domain- and high-fidelity, capable of justifying decisions, validating models and measuring liability. Therefore, explainability in such domains is not a technical option but rather a compliance and accountability requirement, and as such, it leads researchers to consider interpretable-by-design models and more of high-quality evaluation measures.

What is coming out of the analysis is a rising fascination in hybrid XAI methods, whereby the understanding can be incorporated with the deep-learning frameworks. In the example of such models as ProtoPNet that already use prototypical parts instead of individual examples in convolutional architecture, the decision can be justified by the mention of representative examples, an idea based on human intuition. This balanced between accuracy and interpretability, these architectures can be a hopeful compromise between post-hoc explanations and intrinsically interpretable models. Besides, possibilities of human-in-the-loop XAI systems when user feedback is directly used to generate an explanation or optimise the model are highlighted in the literature. These are very useful in adaptive systems such as individual healthcare, recommendation and decision support systems. These methods have their practical difficulties, though, such as interface design, customization on the person, and the usefulness of the explanation. To make these systems better, there should be a need of cross-disciplinary team work of AI developers, human-computer interaction (HCI) specialists, and domain experts to provide explanation by not only by being technically correct but also cognitively meaningful and hence actionable.

Table 5: Summary of Key Findings from Comparative XAI Literature Review

Focus Area	Key Observations	Advantages	Challenges / Limitations
Post-hoc Techniques (e.g., SHAP, LIME)	Widely used due to model-agnostic flexibility and feature attribution capabilities	Applicable to black-box models; useful for local interpretability	May lack fidelity; sensitive to input perturbations; computational overhead in real-time use
Regulatory Domains (Healthcare, Finance)	High demand for interpretable AI due to legal and ethical requirements	Aligns with clinical or financial reasoning; enables auditability	Requires domain-specific explanation formats; strict regulatory compliance
Hybrid XAI Models (e.g., ProtoPNet)	Combine deep learning with embedded interpretability using prototypes or attention mechanisms	Balance between performance and transparency; intuitive explanation formats	Complexity in training; architecture-specific design
Human-in-the-Loop Systems	Leverage user feedback for explanation refinement and adaptive learning	Supports personalization; improves trust and engagement	Requires HCI design; explanation utility varies by user expertise

Conclusion

Explainable Artificial Intelligence (XAI) has appeared as the tenet behind building transparent and accountable and ethically acceptable AI systems, especially with machine learning penetrating health, finance, criminal justice, and autonomous systems. This paper has approached the systematic investigation of the taxonomy of the XAI methods by evaluating the post hoc and intrinsic techniques to analyze their performance and efficiency regarding the most relevant aspects such as interpretability, fidelity, scalability and domain applicability. Although a lot of progress was made, particularly in making model-independent tools such as SHAP and LIME, and incorporating them in hybrid architectures such as ProtoPNet, serious challenges are yet to be overcome. The most prominent ones include lack of common evaluation metrics, the computational hassle of generating explanations and the mismatch between technical explanations and the understanding of users. Moreover, the interpretability is subjective, so it is also difficult to establish universally applicable explanatory systems, and it is necessary to continue to adhere to human-centric points of view and apply domestic horizons. Future research should therefore shift towards interventions in how XAI can be integrated with causal reasoning principles and the human-in-the-loop design to include Theusage of adaptive interfaces to meet individual user profiles and even the cognitive capacity. Interdisciplinary collaboration in order to create policy-aware, context-sensitive, and regulatory-compliant XAI systems is also urgent. Finally, the realization of AI, which should not only be precise but also easy to explain, is needed to advance trust in AI, enlightened decision-making, and ethics of AI application to the society.

References

1. Alvarez-Melis, D., &Jaakkola, T. S. (2018). On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*.
2. Castiñeira, M., & Francis, K. (2025). Model-driven design approaches for embedded systems development: A case study. SCCTS Journal of Embedded Systems Design and Applications, 2(2), 30–38.
3. Chen, H., Li, O., Tao, D., Barnett, A., Su, J., &Rudin, C. (2019). This looks like that: Deep learning for interpretable image recognition. In Advances in Neural Information Processing Systems (NeurIPS).
4. Choi, S.-J., Jang, D.-H., &Jeon, M.-J. (2025). Challenges and opportunities navigation in reconfigurable computing in smart grids. SCCTS Transactions on Reconfigurable Computing, 2(3), 8–17. https://doi.org/10.31838/RCC/02.03.02
5. Doshi-Velez, A., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
6. Jain, S., & Wallace, B. C. (2019). Attention is not explanation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT).
7. Kim, B., Rudin, C., & Shah, J. (2014). The Bayesian Case Model: A generative approach for case-based reasoning and prototype classification. In Advances in Neural Information Processing Systems (NIPS).
8. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (NeurIPS).
9. Middlestadt, P., Russell, C., Wachter, S., &Floridi, L. (2019). Explaining explanations in AI. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT).
10. Muyanja, A., Nabende, P., Okunzi, J., &Kagarura, M. (2025). Metamaterials for revolutionizing modern applications and metasurfaces. Progress in Electronics and Communication Engineering, 2(2), 21–30. https://doi.org/10.31838/PECE/02.02.03
11. Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81–106.
12. Rangisetti, R., & Annapurna, K. (2021). Routing attacks in VANETs. International Journal of Communication and Computer Technologies, 9(2), 1–5.
13. 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, 1135–1144. https://doi.org/10.1145/2939672.2939778
14. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1, 206–215. https://doi.org/10.1038/s42256-019-0048-x
15. 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 (ICCV), 618–626. https://doi.org/10.1109/ICCV.2017.74
16. Tsai, X., & Jing, L. (2025). Hardware-based security for embedded systems: Protection against modern threats. Journal of Integrated VLSI, Embedded and Computing Technologies, 2(2), 9–17. https://doi.org/10.31838/JIVCT/02.02.02
17. Weiwei, L., Xiu, W., &Yifan, J. Z. (2025). Wireless sensor network energy harvesting for IoT applications: Emerging trends. Journal of Wireless Sensor Networks and IoT, 2(1), 50–61