International Journal of
Technology and Emerging Research
ISSN: 3068-109X (online) | ISSN: 3068-1995 (print) | https://www.ioro.org/ijter

International Conference on Intelligent Technologies for Tomorrow (ICITT-2026) · 18 August 2026

Explainable Artificial Intelligence (XAI): Techniques, Applications, Challenges and Future Directions – A Review

Afna Ashraff M [1] , Archana K [2], Buthaina [3] , Krishnaja Radhakrishnan [4] ,Lakshmi K R [5]

[1,2,3,4,5] *3rdBSc Computer Science, Department of Computer Science and Applications, Little Flower College (Autonomous), Guruvayur, Kerala, India*

Emails:afnaashraf2545@gmail.com, krishnajakichu1201@gmail.com, lakshmikr938@gmail.com, achananeeranjanam5@gmail.com, buthainabuthu@gmail.com

Abstract

Machine learning models, and deep neural networks in particular, now inform decisions in healthcare, finance, criminal justice, and other high-stakes domains, yet their internal reasoning remains largely opaque to the people who rely on their outputs. Explainable Artificial Intelligence (XAI) is the body of methods and design principles that render such black-box systems interpretable to developers, regulators, and end users without materially degrading predictive performance. This review surveys the XAI landscape along four dimensions: the major families of explanation techniques—intrinsically interpretable models, post-hoc local methods such as LIME and SHAP, gradient- and attention-based visual explanations such as Grad-CAM, and counterfactual explanations; their realworld applications in domains including healthcare diagnostics, credit and financial risk scoring, and autonomous and safety-critical systems; the challenges that continue to limit adoption, including explanation fidelity, the absence of standardised evaluation metrics, computational cost, and scalability to large generative models; and the future directions the field must pursue. Drawing on a structured review of the recent literature, we compare techniques along scope, fidelity, and cost, and examine the regulatory pressures, including the EU AI Act and the GDPR "right to explanation," that are accelerating adoption. The synthesis shows that no single XAI technique is universally superior; effective explainability requires matching the method to the model class, the application domain, and the stakes of the decision. We conclude that explainability is a necessary, though not sufficient, condition for trustworthy AI, and outline concrete directions for future research, including standardised benchmarks, human-centred evaluation, and explainability for large generative models.

Keywords: explainable AI; interpretability; XAI applications; model transparency; trustworthy AI

1. Introduction

Machine learning models, and deep neural networks in particular, have moved from research benchmarks into decisions that materially affect people's lives: approving a loan, flagging a medical scan, screening a résumé, or recommending a sentence. As these models have grown in capacity, they have also grown in opacity. A gradient-boosted ensemble with thousands of trees or a transformer with billions of parameters cannot be inspected the way a linear regression coefficient table can; its reasoning is distributed across millions of weighted connections that resist direct human interpretation [6], [7].

This opacity is not merely an academic inconvenience. When a system's decision cannot be explained, it becomes difficult to detect bias, diagnose failure modes, satisfy legal obligations, or earn the trust of the people affected by its output. Regulators have begun to respond: the GDPR's provisions on automated decision-making have been read as implying a limited right to explanation [3], and the EU Artificial Intelligence Act imposes transparency obligations on high-risk AI systems [14]. Explainable Artificial

Intelligence (XAI) has emerged as the research area tasked with closing this gap—developing techniques that expose, approximate, or constrain a model's reasoning in terms a human can evaluate.

Despite rapid growth in the number of proposed methods, the field remains fragmented. Terminology is inconsistent across papers, evaluation practices vary widely, and practitioners are often left to choose among techniques—LIME, SHAP, Grad-CAM, counterfactual explanations, and intrinsically interpretable models—without a clear framework for matching a method to a problem [7], [10].

This paper addresses that gap by providing a structured review of explainable AI along four themes signalled in the title. Specifically, the objectives of this review are to: (i) organise the major XAI techniques along consistent dimensions of scope, fidelity, and cost; (ii) survey how these techniques are applied in practice across healthcare, finance, and autonomous and safety-critical systems; (iii) identify the challenges that continue to limit trustworthy, standardised adoption of XAI; and (iv) set out concrete future directions for the field. The remainder of the paper is organised as follows: Section 2 reviews XAI techniques; Section 3 describes the review methodology; Section 4 surveys applications of XAI across real-world domains; Section 5 discusses the challenges these techniques still face; and Section 6 concludes with future directions.

2. Literature Review

Explainability research is commonly organised around two questions: whether a model is interpretable by design, and whether an explanation is local (covering a single prediction) or global (covering the model's behaviour as a whole) [7], [10]. The subsections below synthesise the literature thematically along these lines.

2.1. Intrinsic Interpretability versus Post-Hoc Explanation

Intrinsically interpretable models—linear and logistic regression, decision trees, rule lists, and generalised additive models—are structured so that their internal parameters are themselves a faithful explanation of their behaviour. Rudin [9] argues that for high-stakes decisions, practitioners should prefer such models over post-hoc explanations of a black box whenever an interpretable model can achieve comparable accuracy, since a post-hoc explanation is only an approximation of the true decision process and may not be faithful to it. Post-hoc methods, by contrast, are applied after a black-box model has already been trained, and aim to approximate or probe its behaviour without altering it [7]. A comprehensive practical treatment spanning both families of methods is provided by Molnar [13].

2.2. Local Surrogate and Attribution Methods

The most widely adopted post-hoc techniques are local surrogate and feature-attribution methods. LIME (Local Interpretable Model-agnostic Explanations) perturbs an input, observes the resulting changes in the model's output, and fits a simple, interpretable model—typically sparse linear regression—around the neighbourhood of the instance being explained [1]. SHAP (SHapley Additive exPlanations) draws on cooperative game theory, treating each feature as a "player" and computing its Shapley value as a fair attribution of the prediction across all possible feature coalitions; this gives SHAP a small set of theoretical

guarantees—local accuracy, missingness, and consistency—that LIME does not provide by construction [4]. Both methods are model-agnostic, meaning they can be applied to any predictive model, but this generality comes at the cost of higher computational demand, particularly for SHAP on models with many features.

2.3. Visual and Counterfactual Explanations

For convolutional neural networks used in computer vision, gradient-based visual explanation methods such as Grad-CAM (Gradient-weighted Class Activation Mapping) use the gradients flowing into the final convolutional layer to produce a coarse localisation map that highlights the regions of an image most influential to a prediction [5]. Because Grad-CAM is architecture-specific, it typically executes far faster than model-agnostic perturbation methods, but it is not applicable outside the class of gradient-based, layered visual models it was designed for.

A complementary approach is the counterfactual explanation, which instead of attributing importance to existing features, describes the smallest change to an input that would have produced a different outcome—for example, the minimum income increase that would flip a loan decision from rejected to approved. Wachter et al. [8] proposed counterfactual explanations specifically as a way to satisfy transparency obligations under the GDPR without requiring access to a model's internal logic, since a counterfactual can be generated from input–output behaviour alone.

2.4. Regulatory and Ethical Drivers

Beyond the technical literature, legal and ethical scholarship has been an important driver of XAI adoption. Goodman and Flaxman [3] were among the first to argue that EU data protection law creates a de facto right to explanation for individuals subject to automated decisions. More recently, the EU AI Act formalises transparency and documentation requirements for AI systems classified as high-risk, directly affecting how such systems must be designed and audited [14]. This regulatory pressure has shifted explainability from a desirable research property to, in many sectors, a compliance requirement. More recently, empirical work has begun to benchmark specific model-agnostic XAI methods directly against the AI Act's explainability requirements, showing that persistent gaps remain between what current techniques offer and what the law expects [16].

2.5. Comparative Overview of XAI Techniques

Table 1 summarises the classification of the major XAI technique families reviewed above along four dimensions drawn from the survey literature: scope (local versus global), model dependency (model-agnostic versus model-specific), output form, and relative computational cost [7], [10].

Table 1. Comparison of major explainable-AI technique families.

Technique	Scope	Model dependency	Output form	Relative cost
LIME [1]	Local	Model-agnostic	Sparse linear surrogate	Moderate
SHAP [4]	Local / Global	Model-agnostic	Shapley feature attribution	High
Grad-CAM [5]	Local	Model-specific (CNNs)	Visual saliency map	Low
Counterfactual explanations [8]	Local	Model-agnostic	Minimal input change	Moderate–High
Intrinsically interpretable	Global	N/A (interpretable by	Coefficients / rules	Low

models [9]		design)		

SHAP's guarantees follow from the Shapley value formulation of feature attribution. For a prediction f(x), SHAP expresses the output as a sum of an additive explanation model g, in which each feature i contributes an attribution φ_i, as shown in Eq. (1).

$$g(z') = \varphi_0 + \sum_{i=1}^{M} \varphi_i z'_i$$

Here z′ is a simplified binary representation of feature presence, M is the number of features, φ_0 is the model's expected output, and each φ_i is the Shapley value attributed to feature i; the φ_i sum exactly to the difference between the prediction and the expected output, which is the local accuracy property referenced in Section 2.2 [4].

Figure 1 illustrates the qualitative trade-off between interpretability and predictive performance across common model classes, synthesised from the comparative discussions in [7], [10], [9]. Intrinsically interpretable models such as linear/logistic regression and single decision trees score highest on interpretability but are typically outperformed on complex, high-dimensional tasks by ensemble and deep learning methods, whose internal reasoning is correspondingly harder to inspect.

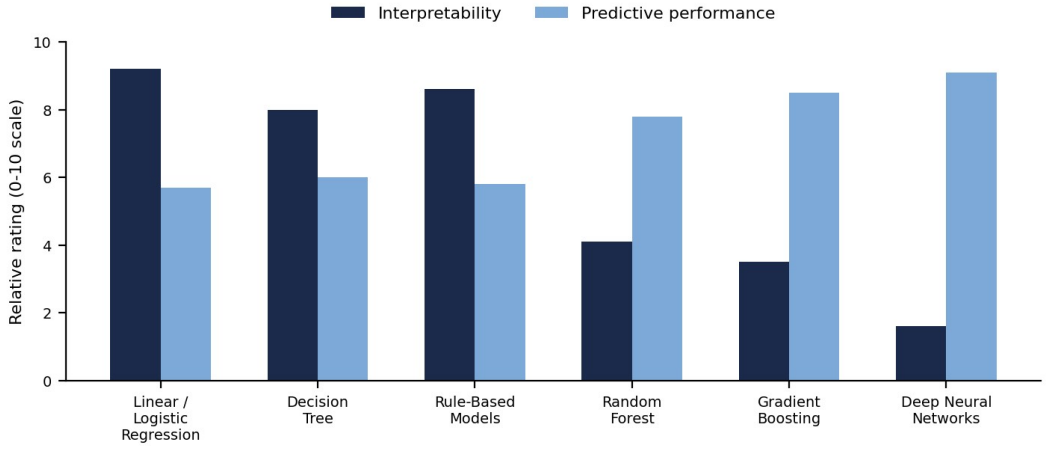

Fig. 1. Qualitative comparison of relative interpretability and predictive performance across common model classes, synthesised from the reviewed literature. Ratings are illustrative, not measured on a benchmark dataset.

3. Methodology

This review follows a structured literature synthesis rather than a formal statistical meta-analysis, which is appropriate given the conceptual and methodological heterogeneity of the XAI field. Sources were identified through the ACM Digital Library, IEEE Xplore, and Google Scholar, using combinations of the search terms "explainable AI," "interpretable machine learning," "model transparency," "feature attribution," and "counterfactual explanation." Foundational papers were included on the basis of citation count and centrality to the field, supplemented by recent survey articles to capture developments not yet reflected in citation metrics.

Although foundational work on interpretable statistical models dates back decades, XAI became a prominent, self-identified subfield of AI research from around 2016 onward, catalysed by the publication of LIME [1] and accelerated after 2018 by the EU General Data Protection Regulation [3] and, more recently, by the entry into force of the EU AI Act in 2024 [14]. This review accordingly concentrates on literature published between

2016 and 2026, with particular attention to the surge of research on explainability for large language and multimodal models that has emerged since 2023 [15].

Each identified technique was then classified and compared along four dimensions drawn from the survey literature [7], [10]: scope (local versus global), model dependency (model-agnostic versus model-specific), output form (feature attribution, visual saliency map, rule set, or counterfactual), and relative computational cost. These dimensions are summarised in Table 1 (Section 2.5) and applied throughout the discussion of applications and challenges below.

4. Applications of XAI Across Domains

XAI techniques are increasingly embedded in real-world decision pipelines. The subsections below illustrate the range of applications reported in the literature, spanning structured-data domains such as healthcare and finance, safety-critical control systems, and the more recent extension of XAI to natural-language and generative systems.

Healthcare. In clinical decision support, SHAP and Grad-CAM are used to justify diagnostic predictions from tabular patient records and medical imaging respectively, helping clinicians verify that a model is attending to clinically relevant features—such as a tumour region in a scan—rather than spurious correlations in the data. Tjoa and Guan [12] survey this “medical XAI” literature and note that explanation is often a precondition for clinical trust and regulatory approval, not merely a diagnostic aid.

Finance. Credit scoring, fraud detection, and algorithmic trading are subject to long-standing regulatory requirements for adverse-action explanations, making XAI a compliance necessity rather than an optional feature. Bussmann et al. [11] apply SHAP to credit risk models to show how individual loan decisions can be attributed to specific financial indicators, supporting both borrower-facing explanations and internal model risk management.

Autonomous and safety-critical systems. In autonomous vehicles and other safety-critical control systems, visual explanation methods such as Grad-CAM are used to audit perception models, verifying that steering or braking decisions are driven by relevant objects in the scene rather than background artefacts. Because failures in these systems carry immediate physical consequences, explanations here are used primarily for pre-deployment verification and post-incident forensic analysis rather than real-time end-user communication [10].

Natural language processing and large language models. As large language models have moved into userfacing applications such as chatbots, summarisation, and automated decision support, explaining their outputs has become a distinct and rapidly growing application area in its own right. Attention visualisation, token-level attribution, and natural-language self-explanations are being adapted from the classificationoriented techniques discussed above to probe why a generative model produced a particular response, though these methods remain less mature and less standardised than their counterparts for smaller, taskspecific models [15].

Criminal justice and public-sector decision-making. Risk-assessment tools used in bail, parole, and sentencing recommendations, as well as automated eligibility screening for public benefits, are high-stakes applications where an explanation is often legally required before a decision can be acted upon. Because these systems affect fundamental rights, XAI here is used less to improve a model’s usability and more to support dueprocess and audit obligations, echoing the regulatory drivers discussed above [10].

5. Challenges of Explainable AI

The technique families reviewed in Section 2, together with the application cases above, point to a set of recurring challenges that continue to limit trustworthy, standardised adoption of XAI:

- Explanation fidelity: post-hoc methods such as LIME approximate rather than expose a model's true decision boundary, so an explanation can be locally plausible yet globally misleading, particularly near decision boundaries or under adversarial perturbation [7], [9].
- Lack of standardised evaluation: there is no consensus metric for explanation quality, making it difficult to compare techniques objectively or certify that a system meets a regulatory transparency requirement [2], [10].
- Computational cost and scalability: exact Shapley value computation is exponential in the number of features, and model-agnostic methods generally scale poorly to the very large feature spaces and parameter counts of modern deep networks [4], [7].
- The accuracy–interpretability trade-off: although Rudin [9] shows this trade-off is not fixed for many structured problems, it remains real for unstructured data such as images and text, where highperforming models are almost always the least transparent.
- Audience mismatch: a single explanation format rarely serves a data scientist debugging a model, a clinician acting on a prediction, and a regulator auditing a system equally well, yet most tools produce one explanation format regardless of audience [10], [12].
- Explainability for large generative models: attribution and surrogate methods developed for classification and regression do not transfer cleanly to large language and multimodal models, leaving a substantial and rapidly growing gap that recent surveys are only beginning to address [15].

These challenges are interconnected rather than independent: the absence of standardised evaluation (Challenge 2) makes it difficult to verify claims of high fidelity (Challenge 1), while computational cost

(Challenge 3) often forces practitioners toward cheaper but less faithful approximations. Addressing them piecemeal is unlikely to be sufficient; the future directions outlined in Section 6 therefore target the underlying gaps in evaluation, tooling, and scope rather than any single technique.

Based on the literature reviewed, we believe that although Explainable Artificial Intelligence (XAI) has significantly improved the transparency of AI systems, several important challenges still limit its widespread adoption. In our view, one of the major challenges is ensuring **explanation fidelity**, as many post-hoc methods provide only approximate explanations that may not accurately represent the model's actual decision-making process. We also observe that the **lack of standardized evaluation metrics** makes it difficult to compare different XAI techniques and assess their reliability across applications. Another challenge is the **high computational cost and scalability**, particularly when applying explanation methods to large and complex deep learning models. Furthermore, achieving the right balance between **model accuracy and interpretability** remains difficult, as highly accurate models are often less transparent. From our perspective, explanations should also be tailored to different stakeholders, since a single explanation may not satisfy the needs of developers, domain experts, end users, and regulatory authorities. Finally, with the rapid growth of **large language models and generative AI**, existing explainability techniques are becoming increasingly inadequate, highlighting the need for more robust, scalable, and human-centered XAI methods. Overall, we believe that addressing these challenges is essential for building trustworthy, transparent, and accountable AI systems in the future.

6. Conclusion and Future Directions

This paper has reviewed Explainable AI along the four themes signalled in its title. It has synthesised the major technique families—intrinsic interpretability, local surrogate and attribution methods, visual saliency methods, and counterfactual explanations—compared them across scope, model dependency, output form, and computational cost, and surveyed their application in healthcare, finance, autonomous systems, natural language processing, and criminal justice and public-sector decision-making. The synthesis shows that explainability is not a single technique to be bolted onto a finished model, but a design consideration that should be weighed alongside accuracy from the outset, particularly in high-stakes and regulated domains. No single method reviewed here is sufficient on its own, and the challenges identified in Section 5 are structural rather than incidental to any one technique.

Based on these challenges, we propose the following priorities for future research:

1. Standardised benchmarks and evaluation protocols for explanation fidelity and human comprehensibility, enabling like-for-like comparison across techniques and domains.
2. Explainability methods designed for large language and multimodal models, extending beyond the classification- and regression-oriented tools surveyed in this review.
3. Human-centred and interactive explanation interfaces that adapt explanation granularity and format to the audience—developer, domain expert, or affected individual—rather than exposing a single fixed output.
4. Robustness of explanations under distributional shift and adversarial manipulation, so that an explanation cannot be gamed independently of the underlying prediction.
5. Tighter integration between XAI tooling and emerging regulatory frameworks such as the EU AI Act, so that compliance can be demonstrated with auditable, reproducible explanations rather than ad hoc reports.

Pursuing these directions would move the field from a loose collection of post-hoc techniques toward explainability as a verifiable, auditable property of AI systems—an outcome that benefits developers, domain experts, regulators, and the individuals ultimately affected by automated decisions alike.

Acknowledgements

The authors thank their department and colleagues for their feedback during the preparation of this manuscript.

Funding

This research received no external funding.

References

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), San Francisco, CA, USA, 2016, pp. 1135–1144.

[2] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.

[3] B. Goodman and S. Flaxman, “European Union regulations on algorithmic decision-making and a ‘right to explanation,’” AI Magazine, vol. 38, no. 3, pp. 50–57, 2017, doi: 10.1609/aimag.v38i3.2741.

[4] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017, pp. 4765–4774.

[5] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in Proc. IEEE Int. Conf. Computer Vision (ICCV), Venice, Italy, 2017, pp. 618–626.

[6] 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, doi: 10.1109/ACCESS.2018.2870052.

[7] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A survey of methods for explaining black box models,” ACM Computing Surveys, vol. 51, no. 5, pp. 1–42, 2018, doi: 10.1145/3236009.

[8] 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.

[9] 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, doi: 10.1038/s42256-019-0048x.

[10] A. B. Arrieta et al., “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” Information Fusion, vol. 58, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.

[11] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, “Explainable AI in fintech risk management,” Frontiers in Artificial Intelligence, vol. 3, art. 26, 2020, doi: 10.3389/frai.2020.00026.

[12] E. Tjoa and C. Guan, “A survey on explainable artificial intelligence (XAI): Toward medical XAI,” IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 11, pp. 4793–4813, 2021, doi:

[13] 1109/TNNLS.2020.3027314.

[14] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2nd ed. Munich, Germany: Christoph Molnar, 2022.

[15] European Parliament and Council of the European Union, Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), Official Journal of the European Union, 2024. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2024/1689/oj. [Accessed: Jul. 20, 2026].

[16] A. Palikhe, Z. Yu, Z. Wang, and W. Zhang, “Towards Transparent AI: A Survey on Explainable Large Language Models,” arXiv preprint arXiv:2506.21812, 2025.

[17] F. Sovrano, G. Vilone, and M. Lognoul, “Assessing Model-Agnostic XAI Methods against EU AI Act Explainability Requirements,” arXiv preprint arXiv:2604.09628, 2026.

Appendix A. Search Strategy

A.1. Search terms: “explainable AI,” “interpretable machine learning,” “model transparency,” “feature attribution,” “counterfactual explanation,” combined with Boolean AND/OR operators across the ACM Digital Library, IEEE Xplore, and Google Scholar.

A.2. Inclusion criteria: peer-reviewed venue or widely cited preprint; direct relevance to explanation methods, evaluation, or regulation for machine learning models; English language.