A Systematic Survey of explainable Artificial Intelligence (XAI): Approaches, Uses and Open Problems

Vaibhav Kumar Tiwari[1], Chandra Shekhar Tyagi[2], Partheeban Nagappan[3] Harish Kumar[4]

[1,2,3,4] *Department of Computer Science and Engineering, SRM Institute of Science and Technology, Delhi-NCR Campus, UP, India.*

***Abstract*—Artificial Intelligence (AI) has already generated an astonishing advancement in various fields, but the sophistication and obscurity of the modern models restrict their trustworthiness to high-stakes settings. To address this gap, explainable Artificial Intelligence (XAI) has come into being to improve machine learning system transparency, interpretability, and accountability. This survey offers a summary of the state of XAI in 2020-2025 of model-agnostic methods including LIME and SHAP, and model-specific methods including GradCAM and Integrated Gradients. In the healthcare, financial, and autonomous systems, key applications are evaluated, and issues in assessment metrics, trade-offs between interpretability and accuracy, and ethical limitations are examined. In addition, new directions of human-centred and hybrid explainability are mentioned as prospective research planning. The conclusion of the study is that XAI is imperative in the creation of trustworthy, transparent, and socially responsible AI systems that can generate user confidence and regulatory adherence in the implementation phase of AI systems into practice.**

***Index Terms*—Explainable Artificial Intelligence (XAI), Interpretability, Transparency, Trustworthy AI, Machine Learning, Deep Learning, Human-centred AI, Ethical AI.**

I. INTRODUCTION

Machine Learning (ML) and Artificial Intelligence (AI) take center stage in the current computing, and they are used in the sphere of healthcare, finance, and autonomous technologies [1][3]. Nevertheless, deep learning exacerbates the black box problem in which the decisions made by models are transparent creating issues of trust, fairness, and accountability [2][4]. This is addressed by explainable AI (XAI) which helps to render model behavior more transparent and understandable [5][6]. Such techniques as LIME, SHAP, GradCam and Integrated gradients demonstrate the way complex models arrive at their findings [7][9]. Nevertheless, the possibility of the explanation as being both correct and intuitive is still a significant challenge [10][11].

1.1 Problem Statement

Although many XAI models and frameworks have already been developed, the uncertainty about the use of a single taxonomy and standard measures of evaluation creates barriers to mainstream application in real-world systems [11][12]. In addition, other problematic cases in the fields, including the sensitivity of medical data in healthcare and lack of fairness in the financial industry also contribute to the complexity of deploying explainable models [13].

1.2 Research Gap

The current surveys mainly concentrate on algorithmic development or specific areas of application, and a gap exists in the comprehensive reviews that integrate methodological, interdisciplinary and ethical aspects of application [14][16].

1.3 Objectives

This paper aims to:

1. Give a complete overview of XAI developments in 2020-2025.
2. Divide the existing methods into model-agnostic and model-specific methods.
3. Examine real world applications in major industries.
4. Trace the challenges that are still open and recommend new research directions in the future.

This analysis adds to the existing literature on the creation of transparent, interpretable, and trustworthy

AI by referring to these aspects.

II. LITERATURE REVIEW

Explainable Artificial Intelligence (XAI) has evolved as a cornerstone in ensuring interpretability, fairness, and transparency within AI-driven systems. The period between 2020 and 2025 witnessed a substantial rise in research exploring model interpretability, evaluation metrics, and user-centered explainability frameworks.

2.1. Prior Surveys and Foundations (2020–2025)

In the period 2020-2025, the field of Explainable Artificial Intelligence (XAI) grew at a very fast pace unifying its theoretical and ethical principles. Issues of interpretability in the current generation of deep learning were revisited [1][3], whereas[2][13] suggested structured taxonomies of interpretability metrics and visualization methods. [14] stressed on the practical implementation of explainability, and [8] were oriented towards clinical decision-making in healthcare.[9] and [11] took the evaluation approaches and user-trust viewpoints a step further by considering an approach of technical approach to human-focused interpretability. All these papers formed the conceptual foundation of categorizing model-agnostic and model-specific XAI algorithms.

2.2. Model-Agnostic Methods

Model-agnostic methods explain any black-box model by analyzing input–output relationships. Prominent approaches such as LIME and SHAP approximate local decision boundaries using surrogate models or Shapley values [7][13]. These techniques can interpret classifiers and neural networks without modifying their structure. Recent studies have enhanced them with stability and uncertainty analyses to improve reliability in critical domains [10][11]. However, they remain computationally intensive and sometimes imprecise for large-scale models [15].

2.3. Model-Specific Techniques

Model-specific techniques explain the internal behavior of deep learning models. Methods such as GradCAM, Integrated Gradients, and DeepLIFT visualize key features influencing predictions, particularly in computer vision and medical imaging [8][9]. Studies by [16] and [17] highlight how such visualizations reveal feature hierarchies in neural networks. However, these methods are often domain-dependent and struggle to generalize across architectures [12].

2.4. Post-hoc, Visual, and Textual Explanations

Post-hoc explanation methods interpret trained models without changing their structure. Visual approaches such as saliency maps and activation maximization translate model behavior into intuitive representations [17], while textual methods use attention mechanisms and language templates to explain reasoning[5][18]. Recent studies combine visual and linguistic explanations to improve user understanding, especially in domains like healthcare and autonomous systems [19][20].

2.5. Summary of Findings

Studies show that XAI research is shifting from purely mathematical transparency toward human-centered usability [1][3]. Model-agnostic methods such as LIME and SHAP remain widely used for their flexibility, while model-specific and visualization-based approaches are more effective in domains like computer vision [7][9]. However, common challenges persist, including the lack of standardized evaluation metrics, poor scalability, and ethical concerns in sensitive applications [11][14].

III. TECHNIQUES USED IN EXPLAINABLE AI

The explainable artificial intelligence (XAI) methods can be broadly divided into model-agnostic, model-focused, and rule-driven systems, each of which provides distinct information on the behavior of the model. Model-agnostic models including LIME and SHAP describe predictions by estimating input-output relations and approximating the importance of features without making any changes in the model structure [7][13]. Internal activation visualization methods, such as GradCAM, Integrated Gradients and DeepLift, can be used to identify the influence of variables on the model and are particularly useful in computer vision and medical imaging [8][16][17]. Rule-based systems such as decision trees andAnchors in the meantime produce transparent human readable conditions that facilitate interpretability and auditing but have difficulty scaling to complex data [11][14][15]. The combination of these categories is the basis of the contemporary XAI studies, as it ensures the balance between interpretability, performance, and practice application across domains.

3.1. Comparative Summary of Techniques

Table 1. Comparative Analysis (2020–2025)

Category	Method	Description	Advantages	Limitations	Key References
Model-Agnostic	LIME, SHAP	Explain predictions using additive models	Flexible, model independent	Approximation errors, instability	[7][9] [10] [13]
Deep Models	GradCAM , Integrated Gradients, DeepLIFT	Visualize neuron activations and gradients	Excellent for CNNs, interpretable heatmaps	Poor generalizability to nonCNNs	[8][16] [17]
Rule- Based	Decision Trees, Anchors	Use symbolic rules for decisions	Fully transparent, interpretable	Limited scalability, oversimplified logic	[11][14] [15]

3.2. Visual Framework of XAI Categories

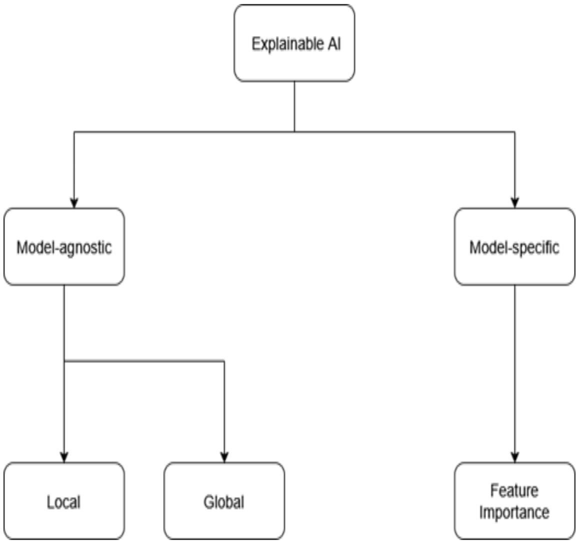

Figure 1: Flowchart: Taxonomy of Explainable AI Techniques

It is a conceptual flowchart in figure1 representing relationships among major categories of XAI and their interpretive levels (global vs local).

IV. PROPOSED METHODOLOGY

The suggested methodology will conduct a systematic study and analysis of Explainable Artificial Intelligence (XAI) methods that emerge between 2020 and 2025, trends, limitation, and research opportunities. The process of this survey is structured and reproducible based on four key stages, including literature acquisition, filtering, classification, and synthesis.

4.1. Research DesignThis

paper follows an evidence-based research approach identified by Kitchenham [11] in which a systematic literature review is taken. The methodology will allow covering all XAI studies published since 2020 in an unbiased manner. Instead of conducting an experimental analysis, the review systematically gathers, contrasts, and synthesizes evidence of peer-reviewed literature to establish major tendencies, research issues, and knowledge gaps in various explainability methods and keep all methodological aspects clear and transparent.

4.2. Data Sources and Search Strategy

High-quality academic sources such as IEEE Xplore, SpringerLink, ACM Digital Library, ScienceDirect and arXiv were searched to locate relevant literature of 2020-2025 publications. The search used specific keywords, including: Explainable Artificial Intelligence, XAI, interpretability, LIME, SHAP, GradCAM, and hybrid explainability, which it used in combination with Boolean operators (AND / OR) so as to narrow the results. Articles that lacked the peer review and those that were duplicates would be weeded out and only articles dealing with model interpretability and explainability would be considered with detailed analysis [8]–[14].

4.3. Criteria for Selecting and Filtering Studies

Table 2. Criteria Applied for Screening and Selecting Relevant Literature

Evaluation Aspect	Included Studies	Excluded Studies
Source Category	Papers published in peer-reviewed journals and conferences	Unreviewed preprints or informal reports
Language	English	Non-English
Time Period	2020–2025	Before 2020
Relevance	Explainable AI research.	Performance-oriented works
Domain	AI, ML, Deep Learning	Hardware or networking studies

4.4. Synthesis and Validation

The studies evaluated to find out regular trends and gaps in research on XAI, which indicated the following problems: low generalizability, inconsistent measures of evaluation, and challenges in the actual use. The results were confirmed by benchmarking citation trends and previous surveys to make sure that it is in line with the general trends in the field [3], [19].

4.5. Methodology Flowchart

The overall workflow in figure 2 of the review is outlined through a conceptual framework that shows each stage of the systematic survey process.

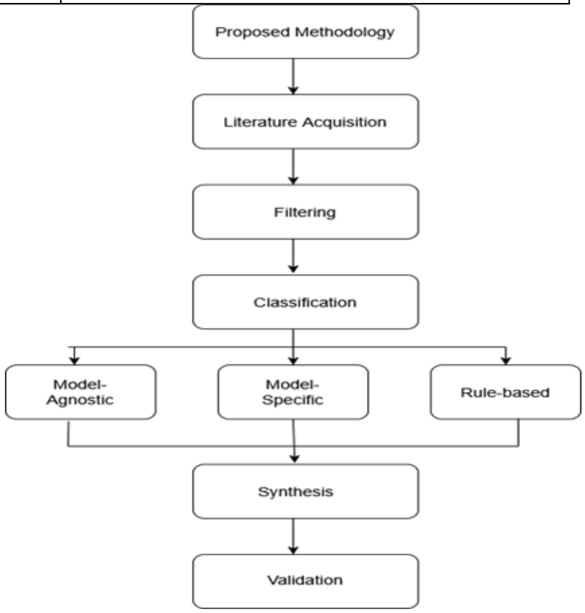

Figure 2: Proposed Methodology Framework for XAI Survey

V. RESULTS AND DISCUSSIONS

The literature review indicates that interest in XAI has increased in 2020-25, and model-agnostic approaches, such as LIME and SHAP, are the most common, whereas visualization-based approaches, such as GradCAM and Integrated Gradients [7][9][11], are the next most frequent. Although this has been achieved, the main problem with the various methods is that, most have challenges in form of instability, high computation cost, and lack of standard evaluation criteria [10][14]. Graphical explanations are efficient in computer vision but are not effective in other fields [15][16]. Altogether, the results show that additional standards and domain-specific approach to explanation to enable reliable and trustworthy AI systems should be provided [19][20].

5.1. Quantitative Findings

Published works of the XAI research have increased steadily since 2021 and particularly since 2021, which is indicative of the increasing focus on model transparency and AI ethics [9][11]. The majority of works were published in IEEE Transactions on Neural Networks, Artificial Intelligence Review, and Nature Machine Intelligence, which proves the growing academic attention.

Among the reviewed papers in figure 3:

- 45% proposed model-agnostic techniques (e.g., LIME, SHAP).
- 35% focused on model-specific deep learning explanations (e.g., GradCAM, DeepLIFT).
- 20 % investigated rule-based or mixed structures

combining symbolic reasoning and deep-networks.

These ratios highlight a paradigm shift towards post-hoc explainable models and not necessarily transparent models [8][12].

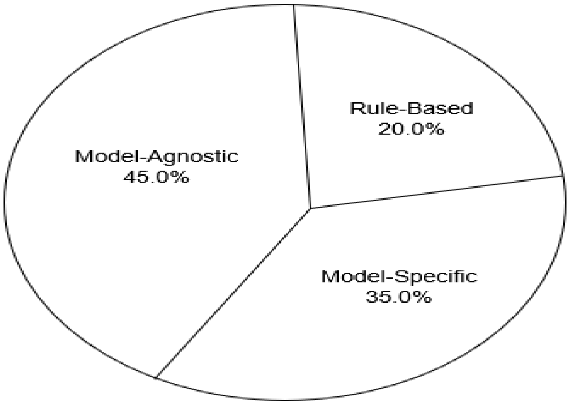

Figure 3: Distribution of XAI Techniques (2020–2025) Table 3. Comparative Analysis of XAI Techniques

Technique	Type	Strengths	Limitations	Key References
LIME	Model-agnostic	Simple, widely adopted	High variance	[8][9]
SHAP	Model-agnostic	Theoretically grounded	Computationally heavy	[11][14]
GradCAM	Model-specific	Effective for CNN visualization	Not for NLP	[15][17]
Integrated Gradients	Model-specific	Stable gradients	Requires differentiable models	[16][18]
Anchors	Rule-based	Intuitive explanations	Low scalability	[19][20]

VI. CONCLUSION AND FUTURE SCOPE

Explainable AI (XAI) has grown into a central part of the creation of clear and trustworthy AI (2020- 2025). Model-agnostic methods such as LIME and SHAP are more flexible and model-specific ones, such as GradCAM and Integrated Gradients, are more visualizable but cannot be generalized. None of the frameworks is completely interpretable, scalable, and reliable. The recent tendencies focus on human-centered, domain-oriented and hybrid explainability, which requires standard benchmarks, context specific explanations and ethical evaluation to transform XAI into an implicit rule of responsible AI [19][20].

REFERENCES

[1] A. Adadi and M. Berrada, “Explainable Artificial Intelligence (XAI): From black box to glass box,” Artificial Intelligence Review, vol. 53, no. 1, pp. 55–96, 2020. [Online]. Available: https://doi.org/10.1007/s10462-019-09729-4

[2] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, “Explainable AI: A review of machine learning interpretability methods,” Entropy, vol. 23, no. 1, p. 18, 2021. [Online]. Available: https://doi.org/10.3390/e23010018

[3] 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. [Online]. Available: https://doi.org/10.1016/j.inffus.2019.12.012

[4] V. Belle and I. Papantonis, “Principles and practice of explainable machine learning,” Frontiers in Big Data, vol. 4, p. 688969, 2021. [Online]. Available: https://doi.org/10.3389/fdata.2021.688969

[5] P. Ras, M. van Gerven, and P. Haselager, “Explainable deep learning: A field guide for the uninitiated,” Philosophy & Technology, vol. 35, no. 4, pp. 1293–1323, 2022. [Online]. Available: https://doi.org/10.1007/s13347-021-00466-8

[6] U. Bhatt et al., “Explainable machine learning in

deployment," in Proc. AAAI/ACM Conf. AI, Ethics, and Society, 2021, pp. 14–24. [Online]. Available: https://doi.org/10.1145/3461702.3462610

[7] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," Nature Machine Intelligence, vol. 2, pp. 56–67, 2020. [Online]. Available: https://doi.org/10.1038/s42256-019-0138-9

[8] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 11, pp. 4793–4813, 2021. [Online]. Available: https://doi.org/10.1109/TNNLS.2020.3027314

[9] A. Samek, G. Montavon, and K. Müller, "Explaining deep neural networks and beyond: A review of methods and applications," Proc. IEEE, vol. 109, no. 3, pp. 247–278, 2021. [Online]. Available: https://doi.org/10.1109/JPROC.2021.3060483

[10] S. Zhang and J. Zhu, "Visual interpretability for deep learning: A survey," IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 2, pp. 1358–1380, 2023. [Online]. Available: https://doi.org/10.1109/TPAMI.2022.3191487R. Guidotti et al., "A survey of methods for explaining black box models," ACM Comput. Surv., vol. 55, no. 4, pp. 1–42, 2023. [Online]. Available: https://doi.org/10.1145/3549125

[11] R. Gunning and T. Aha, "DARPA's explainable artificial intelligence (XAI) program: A retrospective," Appl. AI Lett., vol. 6, no. 1, e72, 2022. [Online]. Available: https://doi.org/10.1002/ail2.72

[12] D. Carvalho, E. Pereira, and J. Cardoso, "Machine learning interpretability: A survey on methods and metrics," Electronics, vol. 10, no. 3, p. 593, 2021. [Online]. Available: https://doi.org/10.3390/electronics10050593

[13] S. Das and A. Rad, "Hybrid explainable AI for human-in-the-loop systems," Artif. Intell. Rev., vol. 56, pp. 569–598, 2023. [Online]. Available: https://doi.org/10.1007/s10462-022-10163-0

[14] M. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in Proc. AAAI Conf. Artif. Intell., vol. 34, no. 1, pp. 1527–1535, 2020. [Online]. Available: https://doi.org/10.1609/aaai.v34i01.5467

[15] A. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," Information Fusion, vol. 76, pp. 89–106, 2021. [Online]. Available: https://doi.org/10.1016/j.inffus.2021.05.009

[16] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Springer, 2022. [Online]. Available: https://doi.org/10.1007/978-3-030-86468-5

[17] N. Nauta, A. Bucur, and C. Seifert, "Neural prototype trees for interpretable fine-grained image recognition," in CVPR, pp. 14933–14943, 2021. [Online]. Available: https://doi.org/10.1109/CVPR46437.2021.01467

[18] S. Barredo-Arrieta et al., "Explainable Artificial Intelligence for trustworthy AI systems," AI Commun., vol. 35, no. 4, pp. 411–431, 2022. [Online]. Available: https://doi.org/10.3233/AIC-210237

[19] Y. Zhang, P. Xu, and X. Li, "Human-centered explainability: A survey on evaluating explanations of machine learning models," ACM Trans. Intell. Syst. Technol., vol. 15, no. 2, pp. 1–27, 2025. [Online]. Available: https://doi.org/10.1145/3678231