Grenze International Journal of Engineering and Technology, January Issue

Measuring the "Good" in XAI: A Critical Examination of Explanation Evaluation Metrics

Kailash C Kandpal[1] and Dr. Prabhat Verma[2]
[1]Department of Computer Science, Harcourt Butler Technical University, Kanpur, India
[2]Dr. Virendra Swarup Institute of Computer Studies

***Abstract*— Explainable AI (XAI) aims to make complex AI models more understandable and trustworthy. However, effectively evaluating the quality of explanations remains a significant challenge. This paper critically examines existing evaluation metrics for XAI, highlighting their limitations and identifying key areas for improvement. We discuss the importance of considering various factors, including fidelity, sparsity, comprehensibility, plausibility, user trust, and the impact of explanations on decision-making. We also emphasize the need for context-specific evaluation and the integration of human-centered perspectives. By addressing these challenges and developing more robust evaluation methodologies, we can ensure that XAI techniques are effective, trustworthy, and beneficial for users.**

***Index Terms*— artificial intelligence, evaluation metrics, explainability, user trust.**

I. INTRODUCTION

Imagine a world where AI systems make life-changing decisions like diagnosing diseases, determining loan eligibility, or even driving a car without anyone truly understanding how they arrived at the conclusions. Due to AI's continues development and integration into our daily lives, this lack of transparency can cause serious risks. It is well known that AI has the power to revolutionize industries, but its black-box nature raises questions about accountability, trust, and fairness. If we don't comprehend the process of making choices of a system, how can we believe it? This lack of interpretability has given birth to a growing movement in the field of Explainable AI (XAI), which aims to make AI models more transparent, interpretable, and trustworthy.

This is where Explainable AI (XAI) becomes important. XAI aims to demystify the 'black box' by providing comprehensible explanations of AI's decision-making processes. By enhancing transparency of the models' conclusions, XAI not only generate trust, but also facilitates accountability and ethical decision-making, particularly in critical sectors like healthcare, finance, and law enforcement. With XAI, we move from blindly trusting AI to actively understanding its reasoning which is further empowering users to make informed decisions and ensuring the technology aligns with societal values. One best example of XAI is, Zest AI [1][2], a company that uses AI to assist lenders in making more accurate credit decisions, by providing explanations for its credit scoring decisions. When a loan application is approved or rejected, Zest AI's platform provides an explanation to the consumer detailing which factors such as income, credit history, or payment behavior were most influential in the final decision. The AI also contextualizes how the applicant's behavior compares to similar candidates and provides recommendations for enhancing their score.

Grenze ID: 01.GIJET.12.1.512

Another illustrative case is Google Health AI's breast cancer detection [3][4] system, which incorporates explainability features, such as visual attention maps to assist doctors in identifying which areas of the mammogram the AI deemed critical in its diagnosis.

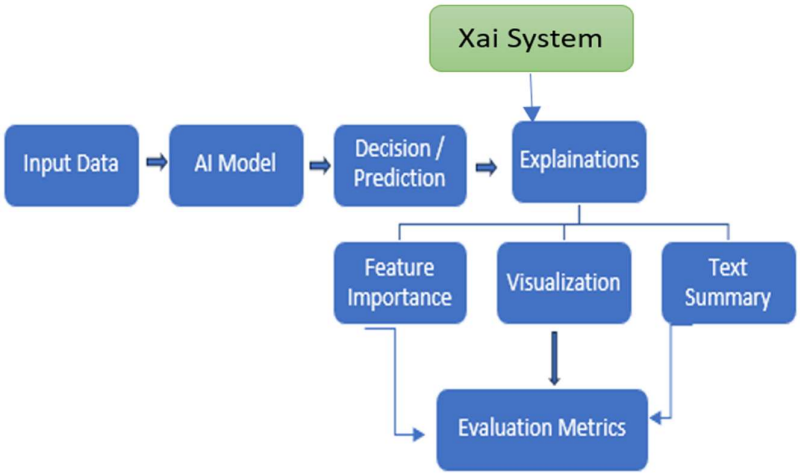

Fig 1. Explanation and evaluation process flow

Nonetheless, delivering explanations constitutes only a portion of the solution. To effectively apply trust and usability, these explanations must be meaningful, precise, and consistent. This leads us to a significant challenge within the domain which further generates the necessity for effective evaluation metrics. To determine whether an explanation actually improves understanding or just makes it more difficult. established measures (like evaluation metrics), to evaluate the effectiveness of XAI methods are required.

This survey paper presents a review and synthesis of insights derived from approximately 20 research studies focused on evaluation metrics in XAI. By evaluating their contributions, the challenges they face, and their proposed resolutions, we seek to offer a thorough overview of the present state of XAI evaluation methods. Additionally, we have investigated the future potential of evaluation metrics in XAI, pinpointing essential areas that require further investigation to guarantee that AI systems are not only effective but also transparent, equitable, and reliable. Furthermore, we propose two innovative qualitative metrics to enhance the evaluation of explanations.

A. The Structure of this Paper

The structure of the paper is organized as follows: Section 2 presents the background of the study, followed by a comprehensive literature review and our metric proposal in Section 3. Section 4 outlines the methodology employed in the research, while Section 5 discusses the future directions for further exploration. Finally, Section 6 offers the concluding remarks of the study.

B. Background

Explainability is broadly divided into local and global explainability (Fig. 2.0). Local explainability focuses on why a model made a specific prediction for a single input. Techniques such as LIME [5] and SHAP [6] provide such instance-level explanations. In contrast, global explainability aims to describe the model's overall behavior across the dataset, revealing general patterns and feature influences. Methods like feature importance [7] and partial dependence plots [8] help identify which features consistently affect model decisions, such as how credit score or income typically influence loan approvals. Explainability can further be classified into intrinsic and post-hoc methods. Intrinsic explainability applies to models that are interpretable by design—e.g., decision trees, linear models, and rule-based classifiers—whose transparent structures make reasoning visible and understandable [9].

Post-hoc explainability, on the other hand, is used for complex, non-transparent models like deep neural networks. It explains predictions after training using tools such as SHAP, LIME, and saliency maps [10], which highlight the input regions that influenced a model's decision—for example, key image areas emphasized during classification [11].

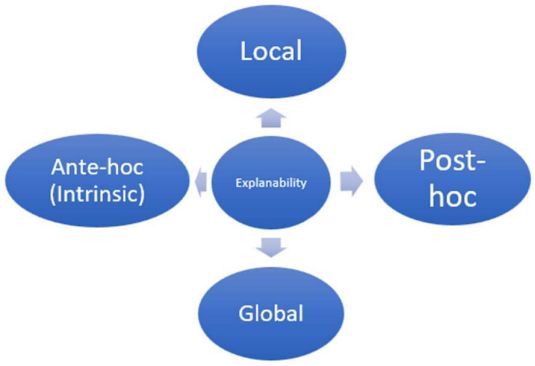

Fig 2. Explanation Categories

Both local/global and intrinsic/post-hoc explainability are vital for interpreting machine learning models. They have different complementary roles as shown in table 1

TABLE 1

Illustrating the complementary roles of explainability approaches in enhancing model interpretability across different use cases and model types.				
Aspect	Local Explainability	Global Explainability	Intrinsic Explainability	Post-Hoc Explainability
Focus	Explains individual predictions	Explains the model's overall behavior	Models are interpretable by design	Explanations provided after model training
Scope	Instance-specific (single prediction)	Dataset-wide (all predictions)	Provides transparency through simple model structures	Explains complex models (e.g., neural networks)
Example Techniques	LIME [5], SHAP [6]	Feature Importance [7], Partial Dependence Plots [8]	Decision Trees, Linear Regression	SHAP, LIME, Saliency Maps [10]
Model Type	Any model (black-box or interpretable)	Any model (black-box or interpretable)	Transparent models (e.g., decision trees, linear regression)	Complex, black-box models (e.g., deep learning)
Use Case	Why did the model make this prediction for this input?	What is the model's general decision-making process?	Easy to understand without external tools	Requires external tools to generate explanations
Granularity	Fine-grained understanding of specific cases	Coarse-grained understanding of model-wide patterns	Directly interpretable from the model itself	Requires external interpretability techniques
Real-World Example	Diagnosing why a patient is classified as high-risk (healthcare)	Understanding general factors influencing loan approvals (finance)	A decision tree model in medical diagnosis	Explaining deep learning model decisions in image analysis

As AI models grow in complexity, particularly in domains like healthcare, finance, and law, ensuring that these models are not only accurate but also transparent and interpretable becomes crucial. Evaluation methods provide structured ways to measure whether the explanations provided by an AI model are understandable, trustworthy, and useful to end users. Evaluation methods are indispensable for ensuring that the explanations generated by AI systems are effective, understandable, and trustworthy. They provide a means to quantify the quality of explanations, enabling comparisons between different techniques and guiding improvements in model design and interpretability. Ultimately, the use of robust evaluation methods ensures that explainable AI fulfills its goal of increasing transparency, trust, and fairness in AI decision-making processes.

II. LITERATURE REVIEW AND METRIC PROPOSAL

In Explainable AI (XAI), evaluation methods can be broadly categorized into qualitative and quantitative approaches. Both are essential for assessing the effectiveness of explanations generated by AI models.

A. Qualitative methods

In Explainable AI (XAI) focus on human-centered evaluations, where the effectiveness of AI explanations is judged based on user experience, expert opinion, and subjective feedback. These methods are particularly valuable when assessing how well explanations align with human intuition, improve trust, and facilitate

decision-making. Doshi-Velez and Kim emphasize the importance of human-grounded evaluation methods [12][36][40], where human subjects assess the interpretability and utility of AI models through user studies. In Think-Aloud Protocol method, participants verbalize their thoughts as they interact with explanations. It provides insights into how users process information and form decisions based on AI explanations. This is valuable in understanding the cognitive load and reasoning patterns influenced by the explanation. Abdul et al. (2018) [13][34] discuss how think-aloud protocols can be used to capture real-time user interactions with AI systems, making them useful for evaluating XAI systems' usability and interpretability. Another qualitative method is focusing group and case studies. Hoffman et al. (2018) [14] highlight the role of case studies in providing contextual evaluations of explainability, especially in complex, high-stakes domains. In expert annotation Domain experts annotate the explanations provided by XAI systems, indicating whether they align with their domain knowledge and reasoning. Miller (2019) [15][39] suggests that expert feedback is essential for assessing the quality of explanations, especially when those explanations are used in critical decision-making scenarios.

B. Quantitative Methods

Quantitative methods involve objective, measurable evaluations of XAI systems. These methods typically rely on mathematical or computational metrics to evaluate how faithfully the explanation represents the model's decision-making process or how useful the explanation is in improving human decision-making. Ribeiro et al. [16][35][38] introduce the concept of fidelity as a core metric in LIME, which evaluates how closely the local explanations approximate the model's predictions. Fidelity refers to how accurately the explanation reflects the underlying model's decision-making process. Higher fidelity means that the explanation is more truthful to the internal workings of the model. Sparsity on the other hand measures the conciseness of the explanation, specifically how many features or components are involved in the explanation. Sparse explanations are easier for humans to understand, as they focus on the most relevant aspects. Lundberg and Lee [17] emphasize the importance of sparsity in SHAP, where fewer features are highlighted to improve interpretability while maintaining explanation quality. Alvarez-Melis and Jaakkola [18] discuss robustness as a key quantitative metric for evaluating XAI, proposing methods to ensure that explanations remain consistent under input perturbations. Surrogate models [16] are often used in XAI techniques (e.g., LIME) to approximate the behavior of more complex models. The accuracy of these surrogate models is an important metric, as it determines how well the surrogate mimics the original model's behavior. Quantitative evaluation can also involve measuring the performance of human decision-makers when provided with explanations. For example, users' accuracy, decision-making speed, and error rates can be compared with and without the use of AI explanations. Lai and Tan [19][33] conducted studies where human performance was evaluated based on the support provided by explanations, showing how explanations can either improve or hinder decision-making. Another one is Perturbation-Based Metrics Montavon et al. [20][31][32]. These metrics assess how much explanations change when the input is perturbed. It helps evaluate the robustness of explanations and ensures they are not overly sensitive to small variations in the input data. Preece et al. [21][30] suggest that combining qualitative and quantitative approaches yields the most reliable assessment of explainability, especially in domains requiring both technical accuracy and user trust.

TABLE II

Provides an overview of various metrics used in both qualitative and quantitative evaluation methods for XAI, highlighting the different approaches and their focus areas				
Category	Metric	Description	Method Type	Application
Qualitative	Satisfaction	User's overall satisfaction with the explanation.	Subjective	User studies, interviews
	Trust	Measures if explanations increase users' trust in the model.	Subjective	User studies, interviews, focus groups
	Comprehensibility	Assesses how easily users understand the explanation.	Subjective	User studies, think-aloud protocols
	Usefulness	Evaluates whether the explanation helps users make decisions.	Subjective	Case studies, focus groups
	Transparency	Users' perception of how transparent the model's decision process is.	Subjective	Expert annotations, user feedback
	Mental Model Fit	Assesses whether the explanation aligns with users' mental model of the problem.	Subjective	Think-aloud protocols, interviews
	Interpretability	Evaluates whether the explanation is simple and understandable.	Subjective	User studies, expert annotations

	Actionability	Determines whether explanations lead users to take informed actions.	Subjective	Focus groups, case studies
	Engagement	Measures the level of interaction and engagement with the explanations.	Subjective	User studies, interviews
	Perceived Fairness	Assesses whether users perceive the explanation as fair.	Subjective	Focus groups, user feedback
Quantitative	Fidelity	Measures how accurately the explanation reflects the model's decision process.	Objective	Local/Global fidelity evaluation
	Sparsity	Measures how concise the explanation is, focusing on fewer key features.	Objective	Explanation sparsity in LIME/SHAP-like methods
	Consistency	Assesses if similar inputs yield similar explanations.	Objective	Consistency analysis across similar samples
	Completeness	Evaluates how much of the decision-making process is covered by the explanation.	Objective	Explanation completeness in feature attribution methods
	Robustness	Measures how stable explanations are when small input perturbations occur.	Objective	Perturbation-based metrics
	Accuracy of Surrogate	Assesses how well surrogate models approximate the original model.	Objective	Surrogate model accuracy (used in LIME-like methods)
	Explanation Accuracy	Compares explanations to ground truth or expert annotations.	Objective	Comparison with expert-labeled ground truth
	Human Performance	Measures improvement in human decision-making with explanations.	Objective	Task-based evaluations of users' decision accuracy with/without explanations
	Computational Efficiency	Evaluates the time or resources needed to generate explanations.	Objective	Algorithmic efficiency, especially in real-time decision systems
	Feature Importance Agreement	Compares feature importance rankings across different explainers.	Objective	Agreement metrics to check consistency of feature importance across methods

C. Inclusion and Exclusion Parameters

As we need most relevant and valuable research articles for this review paper, we developed the parameters through our combined expertise and a broad review of many high quality research works in similar domain.
Major points of selection criteria include:
Inclusion Criteria:

- For ensuring the current research contents papers published in last 7 years are only taken as sample.
- To ensure quality only peer reviewed journals and conferences are considered.

Exclusion Criteria:

- Papers older than 7 years are excluded from study to maintain newness.
- Nonscientific papers, book chapters are excluded to ensure relevancy.
- To ensure unique insights redundancy has been removed.
- Non peer revied content is not included to ensure standardize quality of content.

III. NOVEL METRIC PROPOSAL

In this paper we have proposed 2 qualitative metrics to evaluate the explanations, they are i) Empathy, and ii) Explanatory Bias Detection. Empathy focused on whether the explanation resonates with users' mental models, emotional states, and comfort levels, especially in sensitive decision-making contexts, such as healthcare, finance, or legal systems. It will work with different key qualitative dimensions like

- Emotional sensitivity, for example: In healthcare AI, when informing a patient about a critical diagnosis, the explanation should be presented with a tone of empathy, offering comforting context rather than simply delivering cold facts.
- User-Centered Framing which Focuses on how well the explanation is framed from the user's perspective. Like In a job application AI system, if a candidate is rejected, the explanation should highlight the specific qualifications they need to improve in a constructive way, avoiding blunt or discouraging language.
- Transparency in Uncertainty is another dimension where the AI model has uncertainty or ambiguity in its decision, the explanation should transparently convey this without alarming the user, providing additional

context or suggesting next steps. Like In predictive AI systems (like medical prognosis), when the model is unsure, the explanation should transparently state this but offer possible paths forward (e.g., further tests or consultations) rather than leaving the user uncertain and anxious

- Positive Framing which explores in cases of negative outcomes (e.g., loan denial or job rejection), the explanation should focus on what the user can do to improve their chances in the future. It should be framed in a way that empowers the user rather than simply stating a rejection. Example: For a credit denial decision, the explanation could emphasize, "While your application was not approved this time, improving your credit score by X points could lead to approval in the future."

Explanatory Bias Detection is another metric that we have proposed. As AI systems can inadvertently encode biases, detecting explanatory bias could help ensure that the model is offering fair and neutral explanations, a gap in current evaluation metrics. We could measure this by verifying if explanations are consistent across similar inputs. Inconsistent explanations may indicate bias, as the system's reasoning should remain stable for similar decisions. There are still some challenges in this as It can be difficult to define and measure bias in explanations because different users may perceive the same explanation in different ways, especially when the explanation is intended for a non-expert audience. So it comes up with the future scope also to develop a bias detection metric that measures the degree to which explanations align with fairness principles across different demographic groups or a metric that evaluates how sensitive the explanation is to small changes in input features, ensuring the explanation remains fair and unbiased under different conditions.

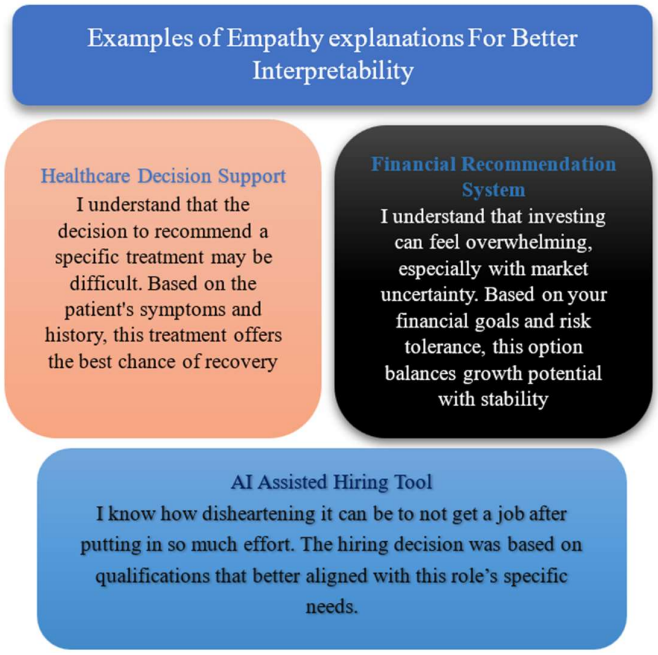

Fig 2. How Empathy metric can help improving interpretability in explanations

IV. METHODOLOGY

This survey paper adopts a systematic approach to analyze and synthesize existing literature on explainability and interpretability in artificial intelligence (AI), with a focus on evaluation metrics, user trust, and challenges in developing interpretable models. The methodology is organized into the following phases:

A. Literature Selection

A comprehensive review of research papers in the field of explainable AI (XAI) was conducted. Relevant papers were identified through databases such as Google Scholar, IEEE Xplore, and arXiv using search terms like "XAI," "machine learning interpretability," "explanation quality metrics," and "trust in AI." The selected papers were published within the last five years, reflecting the current state of research in the field. This review focused on papers discussing various aspects of interpretability, including evaluation frameworks, user studies, trust measurement, and specific challenges across different domains.

B. Categorization of Themes

The reviewed papers were categorized based on key themes emerging from the literature:
Evaluation Metrics: Several papers discussed the lack of consensus on standardized metrics for evaluating interpretability. Papers like [21]and [22] emphasized the need for both qualitative and quantitative metrics that cater to specific domain requirements.
Trust and Usability: Studies such as [23] focused on trust as a critical aspect of user acceptance of AI systems. Trust measurement, explanation usability, and interaction were core themes, especially in sensitive fields like healthcare.
Interdisciplinary Collaboration: The need for collaboration across disciplines was highlighted in papers like [21] and [24] These studies underscore the importance of combining technical and human-centered approaches to enhance understanding and trust in XAI systems.
Ethical and Domain-Specific Considerations: Ethical implications and the need for domain-specific interpretability methods were frequent in papers like [24] The ethical concerns regarding fairness, bias, and the societal impact of AI were prioritized, especially in high-stakes applications like criminal justice and hiring.

C. Analysis of Challenges and Gaps

A thematic analysis of the challenges and limitations identified in the literature was conducted. Common challenges include:
Lack of Standardization: Many papers, including [25] noted the absence of a universally accepted framework for evaluating explanation methods, leading to inconsistencies in research outcomes.
Complexity in Evaluation: Evaluating explanations in machine learning systems is complex and requires a holistic approach. For instance, "Towards Human-Centered Explainable AI" identified that explanations can improve user understanding but may not always translate into increased trust.
Proxy Models and Real-World Application: The challenge of selecting appropriate proxy models for evaluation was highlighted in [29] where trade-offs between interpretability and accuracy were noted as significant issues.

D. Future Research Directions

The future scope discussed in the papers was synthesized to identify key areas for further investigation, including:
Standardized Metrics Development: Several papers, such as [26] pointed to the need for developing domain-specific and standardized metrics for interpretability.
User-Centered Studies: Papers like [25] and "Metrics for Explainable AI: Challenges and Prospects" called for more user-centered approaches, including longitudinal studies to better understand how trust in AI systems evolves over time.
Interactivity and Ethical Considerations: Future work in [24] and [26] emphasized the importance of exploring interactive explanations and integrating ethical considerations, such as fairness and bias, into AI models.

E. Comparative Analysis

Several works emphasize the need to evaluate the quality of explanations produced by XAI systems. Papers [26] and [21] commonly identify fidelity, explanation satisfaction, and trust as central metrics, while [23] adds a usability-driven perspective to explanation quality.
Most papers agree that fidelity—how accurately an explanation reflects the model's actual reasoning—is fundamental. Likewise, user satisfaction appears across both user-centered and system-focused evaluation frameworks. Differences arise in emphasis: works like [25] prioritize user understanding and trust, whereas [23] focuses more on interaction quality and learnability.
Paper [22] introduces appropriate trust, stressing the need to avoid both over-reliance and under-reliance on AI systems. Subjective metrics such as trust, interpretability, and satisfaction are highlighted in [25] and [22], particularly for high-stakes settings like healthcare. In contrast, papers like [23] and [24] advocate more objective criteria—fidelity, stability, and consistency—arguing that such metrics allow reproducible evaluations without requiring users. Still, a key challenge remains: ensuring these objective metrics truly reflect real-world trustworthiness.
A recurring theme is trust, with several works (including "Metrics for Explainable AI" and [24]) arguing that trust is dynamic and must be studied over time. Meanwhile, [22] stresses the importance of calibrated or appropriate trust. Another major issue identified across [25] and [21] is the lack of standardized evaluation frameworks. Without unified metrics, results across XAI methods and domains become difficult to compare, which also undermines reproducibility, as noted in [26]. Finally, many papers—including [25], [23], and [27]—

highlight the importance of interactivity and domain specificity in explanation evaluation. Interactive explanations can enhance understanding and trust, while domain-specific metrics are often essential, as emphasized in [23], [27], and supported by [28], which calls for interdisciplinary collaboration to tailor evaluation protocols to individual application areas.

F. Key Takeaways from the Comparative Analysis

However, some papers, such as [24] caution against an over-reliance on domain-specific methods, advocating instead for a generalized framework that can be adapted but retains core principles like fidelity and intelligibility. Fidelity and Satisfaction are core metrics across various approaches, but how they are measured—whether through subjective user feedback or objective benchmarks—differs widely. Trust is a critical outcome, with an emerging consensus that it needs to be treated as a dynamic process rather than a static quality. There is a gap in standardization, with multiple papers calling for unified frameworks to evaluate XAI across different domains, though there is debate on how flexible or rigid these frameworks should be. Domain-specific needs introduce complexities, and while some argue for tailored metrics, others push for generalized, adaptable frameworks. User interactivity and usability (table 3) are becoming increasingly recognized as vital components for effective explanations, especially in human-AI interaction contexts.

TABLE III

Comparative analysis of Strength and weaknesses of evaluation metrics for Explainable AI (XAI):				
Evaluation Metric	Description	Paper Insights	Strengths	Weaknesses
Fidelity (Faithfulness)	Measures how accurately the explanation reflects the underlying model's decision-making process.	- **SCS (Holzinger et al.)**: Important for assessing model correctness. - **Mohseni et al.**: Emphasizes fidelity when auditing algorithms.	- Ensures accuracy in explanations. - Builds user trust in model behavior.	- Difficult to apply to complex models like deep learning. - High computational cost.
Completeness	Measures the extent to which the explanation covers all relevant factors that influence the model's decision.	- **Salih et al.**: Calls for expert-grounded evaluations in domains like cardiology to ensure completeness. - **Salih et al.**: Proxy-based methods must capture complexity.	- Enhances user understanding of decisions. - Helps in sensitive applications like healthcare.	- May overwhelm users with too much information. - Hard to achieve for complex models.
Consistency	Measures how consistent explanations are for similar inputs.	- **SCS (Holzinger et al.)**: Consistency critical for user trust. - **Salih et al.**: Stable explanations prevent confusion in medical use cases.	- Increases user confidence in the model. - Reduces perceived randomness in AI decisions.	- Might not capture slight variations in decision-making. - Could lack precision in some edge cases.
User Satisfaction	Evaluates how well the explanation aligns with user expectations in terms of clarity and usefulness.	- **Mohseni et al.**: Satisfaction is essential for understanding and decision-making. - **Holzinger et al.**: SCS uses Likert scales to assess satisfaction.	- Ensures user engagement and trust. - Provides insights for improving user experience.	- Subjective nature can lead to varying results. - May be influenced by individual biases.
Trust	Measures how much users trust the explanation and the AI system's decisions.	- **Salih et al.**: Trust is vital in healthcare for AI decisions. - **Mohseni et al.**: User trust influenced by explanation simplicity and transparency.	- Trust enhances system adoption and user interaction. - Builds confidence in high-stakes fields like healthcare.	- Trust can be easily manipulated by overly simplistic explanations. - May require a long time to build.
Actionability	Assesses whether users can take actionable steps based on the explanation provided by the AI.	- **Salih et al.**: Essential in healthcare applications where actionable insights are critical. - **Salih et al.**: Actionability can guide clinical decision-making.	- Directly impacts the usability of XAI in practical settings. - Useful in fields like healthcare, finance.	- Actionability may be limited by model complexity. - Some complex decisions might not translate into actions.
Causal Understanding	Measures how well the explanation conveys causal relationships between features and outcomes.	- **Holzinger et al.**: Identified a gap in causal understanding in current XAI methods. - **Salih et al.**: Calls for causal explanations to align better with human comprehension.	- Improves transparency and accountability. - Increases user understanding of model decisions.	- Difficult to achieve in models that don't explicitly model causality. - Requires domain expertise to interpret.

V. Conclusion

This survey paper has explored the critical role of Explainable AI (XAI) in fostering trust and transparency in the era of increasingly complex AI systems. We have examined a diverse range of research, including techniques for enhancing interpretability, methods for evaluating explanation quality, and the ethical considerations surrounding XAI development.

Our analysis highlights the multifaceted nature of XAI, encompassing various definitions and methods, and the crucial need for user-centered approaches that prioritize understanding and trust. We have identified key challenges in the field, including the lack of standardized evaluation metrics, the subjective nature of human judgment, and the need for context-specific explanations. Ultimately, the goal of XAI is to empower users to understand and interact with AI systems in a meaningful and meaningful way, building trust and enabling responsible AI adoption across various domains.

References

[1] G. Becker et al., "Fairness and Interpretability in AI-Powered Credit Scoring," in Proc. Int. Conf. Fairness, Accountability, and Transparency, 2020.

[2] Zest AI, "Building Better Credit Scoring Models with Explainable AI," Zest AI, 2020. [Online]. Available: https://www.zest.ai

[3] S. M. McKinney et al., "International evaluation of an AI system for breast cancer screening," Nature, vol. 577, no. 7788, pp. 89–94, 2020.

[4] Y. Liu et al., "Artificial intelligence–assisted mammography screening: Evaluation of a deep learning system for detecting breast cancer in mammograms," Radiology, vol. 296, no. 2, pp. 385–394, 2020.

[5] 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, pp. 1135–1144, 2016.

[6] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Adv. Neural Inf. Process. Syst., vol. 30, pp. 4765–4774, 2017.

[7] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[8] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Stat., vol. 29, no. 5, pp. 1189–1232, 2001.

[9] C. Rudin, "Stop explaining black-box machine learning models for high-stakes decisions and use interpretable models instead," Nat. Mach. Intell., vol. 1, no. 5, pp. 206–215, 2019.

[10] [10] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," arXiv preprint arXiv:1312.6034, 2013.

[11] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp. 618–626, 2017.

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

[13] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli, "Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda," in Proc. 2018 CHI Conf. Human Factors Comput. Syst., pp. 1–18, 2018.

[14] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable AI: Challenges and prospects," arXiv preprint arXiv:1812.04608, 2018.

[15] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," Artif. Intell., vol. 267, pp. 1–38, 2019.

[16] 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, pp. 1135–1144, 2016.

[17] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Adv. Neural Inf. Process. Syst., vol. 30, pp. 4765–4774, 2017.

[18] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," arXiv preprint arXiv:1806.08049, 2018.

[19] V. Lai and C. Tan, "On human predictions with explanations and predictions of machine learning models: A case study on deception detection," in Proc. Conf. Fairness, Accountability, and Transparency, pp. 29–38, 2019.

[20] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep Taylor decomposition," Pattern Recognit., vol. 65, pp. 211–222, 2018.

[21] L. Coroama and A. Groza, "Evaluation Metrics in Explainable Artificial Intelligence (XAI)," in Commun. Comput. Inf. Sci., 2022, doi: 10.1007/978-3-031-20319-0_30.

[22] H. Löfström, K. Hammar, and U. Johansson, "A meta survey of quality evaluation criteria in explanation methods," in Lecture Notes Bus. Inf. Process., vol. 452, 2022, doi: 10.1007/978-3-031-07481-3_7.

[23] A. Holzinger, A. Carrington, and H. Müller, "Measuring the quality of explanations: The System Causability Scale (SCS)," KI–Künstliche Intelligenz, vol. 34, no. 2, 2020, doi: 10.1007/s13218-020-00636-z.

[24] Z. C. Lipton, “The mythos of model interpretability,” Commun. ACM, vol. 61, no. 10, 2018, doi: 10.1145/3233231.
[25] Y. Rong et al., “Towards human-centered explainable AI: A survey of user studies for model explanations,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 46, no. 4, 2024, doi: 10.1109/TPAMI.2023.3331846.
[26] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, “Machine learning interpretability: A survey on methods and metrics,” Electronics, vol. 8, no. 8, 2019, doi: 10.3390/electronics8080832.
[27] E. Amparore, A. Perotti, and P. Bajardi, “To trust or not to trust an explanation: Using LEAF to evaluate local linear XAI methods,” PeerJ Comput. Sci., vol. 7, 2021, doi: 10.7717/peerj-cs.479.
[28] M. Nauta et al., “From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI,” ACM Comput. Surv., vol. 55, no. 13, 2023, doi: 10.1145/3583558.
[29] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, “Explaining explainability,” 2018.
[30] L. Rieger and L. K. Hansen, “IROF: A low-resource evaluation metric for explanation methods,” 2020.
[31] G. Plumb, M. Al-Shedivat, A. A. Cabrera, A. Perer, E. Xing, and A. Talwalkar, “Regularizing black-box models for improved interpretability,” 2019.
[32] J. Adebayo et al., “Sanity checks for saliency maps,” 2019.
[33] V. Petsiuk, A. Das, and K. Saenko, “RISE: Randomized input sampling for explanation of black-box models,” 2018.
[34] L. Sixt, M. Granz, and T. Landgraf, “When explanations lie: Why many modified BP attributions fail,” 2019.
[35] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, “A benchmark for interpretability methods in deep neural networks,” 2019.
[36] D. Slack, S. A. Friedler, C. Scheidegger, and C. D. Roy, “Assessing the local interpretability of machine learning models,” 2019.
[37] G. Montavon, W. Samek, and K.-R. Müller, “Methods for interpreting and understanding deep neural networks,” Digit. Signal Process., vol. 73, pp. 1–15, 2017.
[38] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, “Towards better understanding of gradient-based attribution methods for deep neural networks,” 2017.
[39] P. Chalasani, J. Chen, A. R. Chowdhury, S. Jha, and X. Wu, “Concise explanations of neural networks using adversarial training,” 2018.
[40] S. Dasgupta, N. Frost, and M. Moshkovitz, “Framework for evaluating faithfulness of local explanations,” 2020.