Advancing Transparency in Machine Learning: A Technical Review of Explainable AI

Akkala Teja Swaroop[1], Kotha Likhitha[2]
[1]*Student, Department of Computer Science and Engineering NRI Institute of Technology, Vijayawada, India*
[2]*Student, Department of Computer Science and Engineering SIR CR Reddy College of Engineering, Eluru, India*
doi.org/10.64643/IJIRTV12I8-190882-459

***Abstract*—The proliferation of complex machine learning models has accentuated the need for *explainable AI (XAI)* – methods that render black-box models more transparent and understandable. In high-stakes applications (e.g., healthcare diagnosis, financial decisions, autonomous systems), stakeholders demand insights into how models arrive at predictions to ensure trust, fairness, and accountability[1][2]. This paper reviews intrinsic (ante-hoc) and post-hoc interpretability techniques, including *feature-attribution* methods (e.g., LIME, SHAP) and *saliency*-based methods (e.g., Grad-CAM), and contrasts model-specific vs. model-agnostic approaches. We discuss evaluation metrics for XAI — notably *fidelity* (faithfulness), *robustness*, and *human interpretability* — which assess how well explanations reflect the model and how understandable they are to users[3][4]. Real-world applications of XAI in domains such as healthcare, finance, and autonomous vehicles are surveyed, highlighting how explanations support decision-makers and compliance. We address regulatory and ethical implications (e.g. GDPR and forthcoming AI laws) that drive XAI adoption. Finally, we outline recent advances — including the use of large language models for natural-language explanations and mechanistic interpretability — and identify open challenges (e.g., standardizing metrics, balancing accuracy vs. explainability) that must be overcome to realize reliable and user-centered XAI[5][6].**

***Index Terms*—Explainable AI (XAI); model interpretability; LIME; SHAP; Grad-CAM; evaluation metrics; fidelity; robustness; transparency; trust; ethics; healthcare; finance; autonomous systems.**

I. INTRODUCTION

Machine learning (ML) and deep learning models have achieved remarkable performance across diverse tasks, but their *"black-box"* nature hinders human understanding. In critical fields like healthcare, finance, and autonomous systems, the inability to explain model decisions raises concerns about transparency, accountability, and safety[1][7]. For example, a neural network diagnosing disease on medical images may be highly accurate, but physicians are reluctant to trust a system unless it can explain *why* it flagged an anomaly. In finance, regulators demand that automated credit or trading decisions be justifiable to end-users. These concerns prompted the emergence of Explainable AI (XAI), which aims to bridge the gap between complex models and human users by producing explanations or interpretations of model behavior[1][8].

XAI encompasses a spectrum of methods and philosophies. Interpretability is often distinguished from explainability: *interpretability* generally refers to model transparency by design (a transparent model whose internal workings are clear), whereas *explainability* refers to post-hoc techniques that aim to articulate a rationale for a given model's prediction[8]. In other words, an interpretable model (e.g., a simple linear regression) offers intrinsic explanations (its parameters directly show feature effects), while a complex model (e.g., a deep neural network) requires post-hoc analysis to produce a human-understandable explanation. The central goal of XAI is to make ML systems *trustworthy* and *accountable*, enabling users to verify decisions,

detect biases, and comply with ethical and legal requirements[1][2].

Recent literature surveys and reviews underscore this imperative. Cheng *et al.* report that in a corpus of XAI research, domains such as healthcare, autonomous driving, cybersecurity, and finance dominate due to their safety and regulatory stakes[9]. Notably, *healthcare* alone contributed many XAI studies, reflecting the critical need for transparent AI in clinical practice. It is increasingly understood that a *perfect* AI system must not only achieve high accuracy but also provide interpretable, actionable insights to users[10][2]. This paper provides a comprehensive treatment of XAI and model interpretability, organized as follows. Section 2 reviews background concepts and taxonomies of XAI. Section 3 details intrinsic and post-hoc interpretability methods (e.g., LIME, SHAP, Grad-CAM). Section 4 surveys evaluation metrics for XAI. Section 5 highlights real-world applications in healthcare, finance, and autonomous systems. Section 6 discusses regulatory, ethical and societal implications. Section 7 outlines challenges and open research questions. Section 8 considers future directions and recent advancements, and Section 9 concludes.

II. BACKGROUND: CONCEPTS AND TAXONOMIES OF XAI

Interpretability vs Explainability: The terms *interpretability* and *explainability* are often used interchangeably, but can be distinguished. Interpretability refers to the degree to which a human can follow the internal mechanics of a model (e.g., a linear model where feature weights are transparent), whereas explainability usually denotes the ability to generate post-hoc explanations for model outputs[8][11]. In practice, interpretability is sometimes used to describe a model that is inherently transparent (*intrinsic*), while explainability refers to auxiliary techniques applied after training (*extrinsic* or *post-hoc*).

Intrinsic vs Post-hoc Interpretability: Intrinsic (ante-hoc) interpretability involves using models that are transparent by design (e.g., decision trees, linear models, rule-based systems) or augmenting a model's architecture to be interpretable. Such models allow direct inspection of feature relations. For example, a decision tree's structure can be read as a logic path; a linear regression's coefficients indicate feature importance. Post-hoc interpretability, in contrast, treats the trained model as a black box and applies additional analysis to interpret its predictions[12]. Post-hoc methods do not alter the original model's structure but extract explanations via additional algorithms. Common post-hoc approaches include feature attribution (saliency maps, gradient-based methods), example-based explanations (showing similar training instances or counterfactuals), and surrogate models (approximating the black box locally with a simpler model)[12][13].

Model-specific vs Model-agnostic: XAI methods can also be categorized as *model-specific* (designed for a particular class of models) or *model-agnostic* (applicable to any black-box model). For example, Grad-CAM is specific to convolutional neural networks, leveraging gradient maps in CNNs, whereas LIME and SHAP are model-agnostic and can explain any classifier by probing its inputs[14][15]. Model-agnostic methods typically work via perturbation or sampling strategies.

Global vs Local Explanations: Another key distinction is between *global* and *local* interpretability. Global explanations aim to characterize a model's overall behavior (e.g., "Feature X generally increases the prediction"). Local explanations target a single prediction or instance (e.g., "for this patient's case, features A, B, C contributed most to the diagnosis")[16]. Local methods, such as LIME or SHAP values for one input, answer "why" a model made a particular decision. Global methods try to summarize rules or feature trends across all inputs. Both perspectives are valuable: clinicians may need local explanations for individual patient decisions, while regulators may want global guarantees about model fairness.

Explanation Formats: Explanations can also take different forms: feature attributions (numerical importance scores, heatmaps), example-based (nearest neighbor or prototypical examples, counterfactual instances), or textual descriptions (natural-language justifications). Some XAI techniques output visualizations (e.g., saliency maps overlayed on images), while others produce human-readable rules or sentences. Effective XAI often involves combining multiple formats to suit user needs.

III. METHODOLOGIES FOR INTERPRETABILITY

This section details representative XAI techniques, distinguishing intrinsic methods and popular post-hoc methods like LIME, SHAP, and Grad-CAM.

3.1 Intrinsic Interpretability

Intrinsic models incorporate interpretability by design. Common examples include linear models, decision trees, simple rule sets, and other white-box models. These models allow users to see how input features map to outputs directly. For instance, a linear regression's coefficient directly shows the effect of a feature, and decision paths in a tree reveal conditions for each prediction. However, intrinsically interpretable models often trade off accuracy or flexibility; for complex tasks they may underperform deep networks. Hybrid approaches sometimes inject interpretability into complex models (e.g., attention mechanisms or by constraining a model structure) to gain some transparency. Nonetheless, much of modern XAI focuses on explaining inherently opaque (black-box) models rather than reinventing them.

3.2 Post-hoc Interpretability Techniques

Post-hoc methods produce explanations *after* a model is trained. We highlight key techniques:

- LIME (Local Interpretable Model-agnostic Explanations): LIME approximates the black-box model locally with a simple, interpretable surrogate (usually linear) model[14]. Given an instance to explain, LIME generates synthetic perturbations of the input and queries the black-box model to obtain predictions. It then fits a weighted linear model on this synthetic data, weighting points closer to the original instance more heavily. The coefficients of the surrogate indicate the influence of each feature on the prediction. Essentially, LIME answers: "If we only consider small changes around this instance, what simple model best approximates the complex model's behavior?" This yields *local fidelity*: the surrogate should mimic the original model's decision boundary near the instance of interest[14]. A simple formulation of LIME's objective is:

  $$\text{explanation}(\mathbf{x}) = \underset{g \in G}{\operatorname{argmin}}\, \mathcal{L}(f, g, \pi_{\mathbf{x}}) + \Omega(g),$$

  where f is the original model, g is an interpretable model from a class G (e.g. sparse linear), $\mathcal{L}$ measures fidelity to f on data weighted by proximity $\pi_{\mathbf{x}}$, and $\Omega(g)$ penalizes complexity[17]. In practice, LIME often uses Lasso regression as g to force sparsity, selecting a small number of features (K-LASSO variant[18]). LIME is *model-agnostic* (works with any classifier) and produces feature importance scores for the given instance. It has been widely applied in tabular, text, and image domains[14][19].
- SHAP (Shapley Additive Explanations): SHAP values come from cooperative game theory and assign each feature an importance value based on Shapley values[20][21]. The idea is to compute how each feature contributes to the difference between the actual prediction and a baseline by averaging over all permutations of feature inclusion. Lundberg and Lee (2017) formulated SHAP as a unifying framework: SHAP values are the unique additive feature attributions that satisfy desirable properties such as local accuracy and consistency. SHAP can be implemented model-agnostically (by approximating all coalitions) or specially (e.g., TreeSHAP for tree models). It inherently provides local explanations: for a single instance, each feature has a Shapley value indicating its contribution. SHAP tends to be computationally expensive (it requires many model evaluations), but recent algorithms have improved its efficiency for common model types. The theoretical guarantees of Shapley (fairness axioms) make SHAP appealing for giving principled, if not always intuitive, explanations[20][21].
- Gradient-based Saliency (including Grad-CAM and Integrated Gradients): For deep neural networks, explanations often take the form of *attribution heatmaps* highlighting important input regions. Grad-CAM (Gradient-weighted Class Activation Mapping) uses the gradients of a target class score flowing into the last convolutional layers to produce a coarse localization map of important regions[22]. Specifically, Grad-CAM computes the gradient of the class score with respect to feature maps, averages these gradients as importance weights, and forms a heatmap by a weighted combination

of feature maps. This highlights which parts of the input image contributed to the prediction. Integrated Gradients is another popular method: it computes the path integral of gradients from a baseline input (e.g., a black image) to the actual input[23]. This aggregates the gradient information along the straightline path in input space, attributing how each input dimension changes the output cumulatively. Both Grad-CAM and Integrated Gradients rely on the model's own gradients and so are *model-specific* (they require differentiable architectures). They yield intuitive visual or input-space saliency maps, especially for vision models[22][24].

- Other Methods: Many other XAI techniques exist. For example, *counterfactual explanations* generate hypothetical perturbations of an input that would change the prediction, answering "what if" questions. *Example-based explanations* (prototypes or influential training points) identify specific data examples that are most representative or most impacted by a prediction. *Neurons and concepts* can be probed for interpretability (e.g., concept activation vectors, dissection methods) to link internal representations to human concepts. Attention mechanisms in language models can also be viewed as explanatory (indicating which words influenced a decision). Some recent work uses *LLMs (large language models)* themselves to generate narrative explanations of other models' outputs, bridging modalities between model inference and human language[25]. Furthermore, approaches like *mechanistic interpretability* attempt to decode network internals (studying individual neurons or circuits) to achieve transparency, though this is still an emerging research area[26].

Each method has strengths and limitations. Local surrogate methods (LIME, SHAP) are model-agnostic but may lack fidelity if the model's decision boundary is very non-linear near the point. Gradient methods assume continuity and may fail on highly non-linear or non-differentiable components. No single technique universally solves explainability, and often multiple methods are applied in tandem.

IV. EVALUATION METRICS FOR XAI

Evaluating explanations is challenging due to subjectivity, but some formal metrics are used to quantify explanation quality. Important criteria include:

- Fidelity (Faithfulness): This measures how well an explanation reflects the actual reasoning of the original model. A faithful explanation should align with the model's true decision process. One way to quantify fidelity is to measure how closely a surrogate model (from LIME, for example) approximates the black box's predictions locally[3]. Other approaches perturb the input according to the explanation and check if the model's outputs change accordingly. High fidelity means the explanation is consistent with the model's internal logic.
- Robustness: This assesses the stability of explanations under small input or model changes[27]. A robust explanation should not drastically change if the input is slightly perturbed in non-influential ways. Robustness metrics might test how much explanation scores vary when adding noise to the input. Lack of robustness (sensitivity to perturbations) indicates unreliable explanations. Additionally, some works point out that certain explanations can be adversarially manipulated—i.e., one can deliberately produce misleading explanations without altering the model predictions[28].
- Human Interpretability (Complexity): This concerns how understandable or usable explanations are to humans[4]. An explanation with fewer elements or simpler form is generally easier for a human to grasp (analogous to Occam's razor). Metrics here may penalize explanation complexity (e.g., number of features used) or evaluate explanations in user studies. For instance, explanations might be scored by how quickly users can simulate the model's behavior or by subjective satisfaction. While formalizing human interpretability is difficult, a proxy is often the *simplicity* of the explanation (e.g., sparsity of feature importance).

Other proposed metrics include *localization accuracy* (for vision, how well a saliency map overlaps ground-truth regions) and *completeness* (how much

of the model's prediction is accounted for by the explanation). Recent work emphasizes that no single metric suffices: fidelity, robustness, and human-centric criteria must all be considered together[3][4]. In practice, XAI evaluation often combines automated metrics with human subject studies. For example, one may measure fidelity by correlation with model outputs, test robustness via adversarial explanation checks, and conduct user surveys for clarity.

Standardized evaluation remains an open challenge[29]; without agreed benchmarks and metrics, comparing XAI methods is difficult. This motivates ongoing research into formalizing metrics (e.g., causal benchmarks, task-specific scores) that can faithfully predict human trust and utility of explanations.

V. APPLICATIONS OF XAI

5.1 Healthcare

Healthcare is among the most prominent domains for XAI, due to the high stakes of clinical decisions. Deep learning is increasingly used for medical imaging (e.g., tumor detection from MRI), diagnostic classification, and predictive analytics. However, medical practitioners require interpretability to trust AI recommendations. XAI techniques have been applied to highlight image regions influencing a diagnosis (e.g., overlaying saliency or Grad-CAM heatmaps on X-rays) or to extract relevant risk factors from patient data. Surveys note that the majority of recent XAI applications target healthcare tasks[9][30]. In practice, XAI helps clinicians verify AI outputs: for instance, a radiologist can see if the model focused on a tumor region or irrelevant artifact. Intrinsically interpretable models (like scoring systems) are sometimes used in parallel with complex models for cross-checking, and post-hoc explanations (LIME, SHAP) are used to justify individual predictions. The EU and US have stressed that medical AI must provide *"convincing evidence of decisions"*[31], making XAI an indispensable tool for adoption. Despite progress, challenges such as limited data, regulatory constraints, and ensuring patient privacy mean that XAI in healthcare is an active research area.

5.2 Finance

Financial services (banking, insurance, trading) increasingly rely on AI for credit scoring, fraud detection, risk assessment, and algorithmic trading. In these sectors, explainability is crucial for regulatory compliance and customer trust. For example, if an AI model denies a loan, regulators (e.g., under fair lending laws) may require an explanation of the decision factors. XAI methods are used to identify which financial features (income, credit history, etc.) drove a credit decision[32]. Banks also incorporate interpretability into model development: simpler models are favored when possible, and when using complex models, explainable interfaces (dashboards, rule extraction) help analysts understand them[33][34]. Deloitte Insights notes that regulators encourage banks to evaluate model complexity versus interpretability, and even require that customers can request explanations of automated decisions[32][35]. Several financial institutions now have dedicated XAI teams to audit models for fairness and transparency. XAI also aids in detecting bias (e.g., against demographics) and meets legal standards like the EU's transparency requirements[36][32]. The trade-off between predictive power and explainability is a key concern in finance: many institutions establish parallel simpler models (for risk management) alongside black-box models, using XAI to manage this balance[32].

5.3 Autonomous Systems

In autonomous vehicles and robotics, AI systems make real-time safety-critical decisions (object detection, path planning, collision avoidance). Here XAI helps engineers and regulators ensure safety and trust. For example, an autonomous car's perception model might be explained by showing which image regions (pedestrian vs. background) influenced its braking decision. Post-hoc saliency maps (Grad-CAM, etc.) and counterfactuals can be used to analyze failure cases (e.g., "why did the car swerve?"). Real-time explainability (e.g., along with sensor readings) may even be required for debugging. Like other domains, regulations for autonomous vehicles may mandate some transparency; currently, XAI research in this area focuses on creating interpretable perception modules and safety verifications. Overall, XAI applications in

autonomous systems are aimed at validating AI decisions to engineers and the public, reducing risk and accelerating deployment.

5.4 Other Domains

While healthcare, finance, and autonomous vehicles are often highlighted, XAI is also important in domains like criminal justice (risk assessment tools), cybersecurity (interpreting anomaly detectors), and customer service. Any application where AI decisions have legal, ethical, or safety implications benefits from explainability.

VI. REGULATORY AND ETHICAL IMPLICATIONS

The rise of XAI is tightly coupled with regulatory and ethical demands. Data protection and AI regulations increasingly require transparency and accountability for automated decisions. For instance, the EU's General Data Protection Regulation (GDPR) has been interpreted to entail a "right to explanation" for automated individual decisions[36]. The forthcoming EU AI Act explicitly mandates transparency for high-risk AI systems, requiring them to provide clear information on how they operate. In banking, authorities have advised institutions to justify black-box model choices and to allow customers to request explanations[32][35]. Similarly, financial regulators in multiple jurisdictions seek information on how firms ensure model explainability.

Ethically, XAI addresses issues of bias, fairness, and trust. AI models can inadvertently encode societal biases; explainability can help detect and correct such biases[37][38]. For example, if a hiring algorithm favors one demographic group, XAI techniques can reveal the features causing the bias (e.g., proxy variables). Transparent explanations also help users trust AI recommendations and catch errors. In domains like criminal justice, opaque AI could reinforce discrimination; XAI provides a check by making decision factors explicit. Leading ethics guidelines (e.g. from EU High-Level Expert Group on AI) emphasize *human-centric XAI*, recommending that explanations be understandable to affected users and that human oversight (human-in-the-loop) be integrated[39][2].

However, caution is needed: an explainable interface can give an illusion of understanding if the underlying model is flawed. Effective XAI requires careful design to avoid misleading explanations. Privacy is also a concern: some explanation methods (e.g. example-based) could reveal sensitive information about training data. Regulators are beginning to specify standards for XAI evaluation (e.g. demanding robustness tests), but the field is evolving. In summary, regulatory and ethical considerations are a major driver of XAI: laws like GDPR and AI Act, along with corporate ethics frameworks, are making explainability a requirement rather than a nice-to-have. This trend likely means that XAI will be an integral part of responsible AI governance in all sectors.

VII. CHALLENGES AND OPEN ISSUES

Despite progress, XAI faces numerous challenges. Key issues include:

- Accuracy vs Explainability Trade-off: There is often a trade-off between model complexity (and thus accuracy) and interpretability. Simple models are interpretable but may underperform on complex tasks, while highly accurate models (deep nets, ensembles) are opaque[10][5]. Balancing this trade-off requires choosing appropriate contexts for explanations. For example, some tasks may tolerate simpler models, whereas others demand black-box models with strong performance accompanied by post-hoc explanations.
- Evaluation Standardization: As noted, there is no consensus on how to quantify explanation quality. Metrics like fidelity, robustness, and complexity exist, but methods often optimize for one at the expense of others. This fragmentation makes it hard to compare XAI techniques[29]. Moreover, many evaluation studies ignore human factors. The lack of ground truth for explanations means research relies on proxies or limited user studies[3][29]. Developing comprehensive benchmarks for different domains is an open challenge[3][29].
- Model Misbehavior and Robustness: Explanations themselves can be manipulated. Adversarial attacks on XAI methods can produce

misleading explanations even when model outputs are unchanged[40]. Ensuring explanations are robust against such attacks is critical. Additionally, explanations may not generalize across slight input shifts (violating robustness). Ensuring that XAI is reliable under real-world noise is an unresolved challenge[27].

- Human Factors: Even a faithful explanation can fail if it is not presented in a human-friendly way. Different audiences (experts vs. laypersons) need different explanation styles[39][2]. Personalizing explanations to user expertise is complex. Cognitive biases can also affect how users interpret explanations. Designing XAI with human-centered evaluation (e.g., HCI studies) is needed but under-explored.
- Complex Models and Context: Modern AI systems, including multi-modal models and large language models, present new difficulties. Current XAI methods were mostly developed for CNNs or decision trees. As [53] highlights, many metrics and tools neglect advanced models (LLMs, multi-modal networks). New explanations that account for complex architecture (and their scale) must be devised[29][41].
- Domain-specific Challenges: XAI in real applications has domain-specific issues. In medicine, explanations must align with clinical reasoning and integrate medical knowledge. In finance, explanations need to meet regulatory standards. Customizing XAI for context (e.g., accounting for legal definitions of fairness) is challenging.
- Ethical and Security Concerns: Providing explanations also risks revealing proprietary model information or sensitive data. For instance, showing influential training examples might leak personal data. Balancing transparency with privacy/security is an ongoing concern.

In summary, while XAI methods proliferate, their actual utility is hampered by these open issues. Key research directions identified in surveys include: balancing fidelity with simplicity, improving computational efficiency and scalability of explanations, defending against adversarial explanation attacks, integrating domain knowledge into explanations, and developing user-centric interactive XAI systems[5][6].

VIII. RECENT ADVANCEMENTS AND FUTURE DIRECTIONS

Recent research is pushing XAI into new frontiers. A notable trend is leveraging large language models (LLMs) to *generate* or *interpret* explanations. As surveyed by Bilal *et al.*, LLMs can translate complex model behaviors into natural-language narratives, making AI decisions more accessible[42][25]. For example, an image classifier could output both a label and an LLM-crafted explanation ("The model identified a pneumonia pattern because it saw shadows in the lower left lung region consistent with fluid accumulation")[25]. Early work also uses LLMs to answer user queries about model behavior or to summarize feature attributions. This line of work suggests future XAI will be more interactive and multi-modal, blending visual or structured explanations with text dialogue.

Another advancement is in *mechanistic interpretability*. Researchers (e.g., Olah *et al.*) are analyzing neural networks at the neuron and circuit level to identify meaningful components (like "this neuron detects vertical edges" or "this circuit encodes the concept of stripes"). While still nascent, such approaches aim to open the black box by reverse-engineering its internals[26].

Techniques continue to evolve in both accuracy and generality: for example, improvements in SHAP (faster algorithms for tree ensembles), extended LIME (GraphLIME for graph data[43]), and new saliency methods (e.g., SmoothGrad, Score-CAM). Attention mechanisms in transformers are also studied as implicit explanations, though their reliability is debated[44].

In terms of evaluation and deployment, future work should focus on establishing standard benchmarks and user-centric assessments[29][6]. This includes creating labeled datasets for XAI tasks, building trust metrics that predict human satisfaction, and integrating adversarial testing for XAI. Interdisciplinary research drawing from cognitive science, law, and human-computer interaction will be essential to make explanations meaningful in context. For instance, guidelines suggest making explanations intuitive and tailored to the audience[39].

Another promising direction is *explanation-guided model training*, where interpretability considerations are built into the learning process. For example, one can impose that learned features align with known concepts or constrain model complexity for eventual interpretability. Finally, as regulators formalize requirements (e.g., EU AI Act), compliance-driven XAI will emerge, where explanation systems must produce audit trails that satisfy legal audits.

In summary, future XAI research will likely be characterized by (1) improved methods that handle the scale of modern AI (LLMs, vision-language models), (2) human-centered design and evaluation, (3) standardized metrics and benchmarks, and (4) integration with regulatory frameworks. These efforts aim to realize XAI that is not just a technical afterthought, but a core component of trustworthy AI[6][2].

IX. CONCLUSION

Explainable AI has become a critical complement to advanced machine learning, addressing the need for transparency, trust, and accountability in AI-driven decisions. This paper reviewed the conceptual foundations of XAI, surveyed both intrinsic and post-hoc interpretability methods (e.g., linear models, LIME, SHAP, Grad-CAM), and discussed key evaluation criteria such as fidelity, robustness, and human interpretability. We illustrated how XAI is applied in sensitive domains like healthcare, finance, and autonomous systems, where explanations support decision-making and regulatory compliance. Challenges remain in balancing accuracy with interpretability, standardizing evaluation, and ensuring explanations are robust and user-friendly. We noted how recent trends (notably the use of large language models for generating explanations) and ongoing research are expanding XAI's capabilities.

As AI permeates more aspects of society, the demand for explainability will only grow. Collaborative efforts among researchers, industry, and policymakers will be required to develop reliable metrics, ethical guidelines, and technologies that make AI systems not only powerful but also transparent and fair. Only then can we ensure that AI augments human decision-making in a responsible and understandable way[1][2].

REFERENCES

[1] [1] Cheng Z., Wu Y., Li Y., Cai L., Ihnaini B., *"A Comprehensive Review of Explainable AI (XAI) in Computer Vision,"* Sensors, vol.25, no.13, p.4166, 2025. [Online]. Available: https://doi.org/10.3390/s25134166.

[2] [2] Chaddad A., Peng J., Xu J., Bouridane A., *"Survey of Explainable AI Techniques in Healthcare,"* Sensors, vol.23, no.2, p.634, 2023. DOI:10.3390/s23020634.

[3] [3] Nazim S., Alam M. M., Hussain S. S., Moinuddin M., Zubair M., Tanweer R., *"Advancing Malware Imagery Classification with Explainable Deep Learning: using SHAP, LIME and Grad-CAM,"* PLOS ONE, vol.20, no.5, May 2025. DOI: 10.1371/journal.pone.0318542.

[4] [4] Seth P., Sankarapu V. K., *"Bridging the Gap in XAI—The Need for Reliable Metrics in Explainability and Compliance," Proc. of ICML*, July 2025. (ArXiv:2502.04695)

[5] [5] Bilal A., Ebert D., Lin B., *"LLMs for Explainable AI: A Comprehensive Survey,"* ACM Trans. Intell. Sys. Technol., 2025. (ArXiv:2504.00125)

[6] [6] Deloitte Insights, *"Explainable Artificial Intelligence (XAI) in Banking,"* Deloitte Center for Regulatory Strategy, 2023. Available: https://www.deloitte.com/insights.

[7] [1] [9] [22] A Comprehensive Review of Explainable Artificial Intelligence (XAI) in Computer Vision

[8] https://www.mdpi.com/1424-8220/25/13/4166

[9] [2] [6] [7] [25] [38] [41] [42] LLMs for Explainable AI: A Comprehensive Survey

[10] https://arxiv.org/html/2504.00125v1

[11] [3] [4] [23] [24] [26] [27] [28] [29] [40] Bridging the Gap in XAI—The Need for Reliable Metrics in Explainability and Compliance

[12] https://arxiv.org/html/2502.04695v1

[13] [5] [10] Survey on Explainable AI: Techniques, Challenges and Open Issues

[14] https://dmas.lab.mcgill.ca/fung/pub/ALF24eswa_preprint.pdf

[15] [8] [13] [15] [19] [20] [21] Advancing malware imagery classification with explainable deep learning: A state-of-the-art approach using SHAP, LIME and Grad-CAM | PLOS One

[16] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0318542
[17] [11] [12] [14] [16] [17] [18] [30] [31] [43] [44] Survey of Explainable AI Techniques in Healthcare
[18] https://www.mdpi.com/1424-8220/23/2/634
[19] [32] [33] [34] [35] [39] Explainable artificial intelligence (XAI) in banking | Deloitte Insights
[20] https://www.deloitte.com/us/en/insights/industry/financial-services/explainable-ai-in-banking.html
[21] [36] [37] The Rise of Explainable AI (XAI) | Onyx
[22] https://www.onyxgs.com/blog/rise-explainable-ai-xai