

Paper Type: Review Article

A Comprehensive Review of Explainable Artificial Intelligence (XAI): Methods, Challenges, and Future Directions

Manar Abdulkareem Al-Abaji ^{1,*} , Meead Mohammed Salih ¹  and Maher Khalaf Hussein ² 

¹ Department of Computer Science, College of Education for Pure Science, University of Mosul, Nineveh, Iraq;

Emails: manar_alabaji@uomosul.edu.iq; meaad_mahammed@uomosul.edu.iq.

² Department of Computer Science, Faculty of Education, University of Telafer, Nineveh, Iraq;

maher.k.hussein@uotelafer.edu.iq.

Received: 08 Dec 2025

Revised: 31 Mar 2026

Accepted: 14 Jun 2026

Published: 27 Jun 2026

Abstract

The growing use of Deep Learning and other complex machine learning models in safety-critical areas raises fundamental questions of model interpretability and explainability in human understandable terms. Explainable Artificial Intelligence (XAI) has recently emerged as a field of research that aims at making complex machine learning models more transparent, explainable and accountable, and this review provides a comprehensive, structured and formal analysis of the current state of XAI along four dimensions: (1) explanation types (model-agnostic vs. model-specific, global vs. local, intrinsic vs. posthoc, etc.), (2) explanation methods (salience maps, feature importance, model interpretability, surrogate models and model explainability), (3) evaluation metrics (faithfulness, interpretability, stability) and (4) applications (healthcare, computer vision, natural language processing, recommender systems, time series analysis, etc.), while also discussing open research directions towards advancing the state-of-the-art in XAI and providing guidance for readers to navigate the current body of knowledge in this area based on their interests and expertise. This review is based on a survey of more than 40 papers between 2020 and 2026 and is aimed at providing a concise and informative review to researchers and practitioners in the field of XAI.

Keywords: Explainable AI; Interpretability; SHAP; LIME; Attention; Grad-CAM; Trustworthy AI; Post-hoc Explanation.

1 | Introduction

Artificial intelligence (AI) systems, especially those based on deep neural networks, have achieved impressive performance in various tasks, including image recognition [1], natural language processing, medical diagnosis [2], and autonomous driving [3], however, most high-performing models are black boxes: their internal logic is not interpretable to humans, neither the users nor the developers nor regulators. In high-stakes settings, this lack of interpretability is not only a nuisance but a source of risk, because a doctor cannot act on an AI system's diagnosis without understanding why the diagnosis is given. Moreover, a judge cannot assess the fairness of a defendant's sentence without tracing the reasoning behind an algorithmic risk score, and a financial regulator cannot enforce accountability after a trading algorithm makes an illegal decision without knowing why the decision was made.



Corresponding Author: manar_alabaji@uomosul.edu.iq



<https://doi.org/10.61356/j.nois.2026.10658>



Licensee **Neutrosophic Optimization and Intelligent systems**. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0>).

Explainable Artificial Intelligence (XAI) is a body of research concerned with making AI systems transparent, accountable and interpretable. The term XAI was coined with the launch of the DARPA XAI program in 2016, yet the concepts and motivations behind XAI have a long and rich history, dating back to early work on rule-based expert systems, case-based reasoning, and inference in Bayesian networks [4]. What makes modern XAI so compelling is the unprecedented scope, diversity, and urgency of the problem, because the European Union's General Data Protection Regulation (GDPR) and AI Act, which enshrine a "right to explanation" for automated decisions, have converted explainability from a largely academic concern into a regulatory requirement. Thus, organizations are now really forced to prove that their models are not discriminatory, that the decision-making process is transparent, and that people have real ways to question and appeal AI predictions [5]. The literature on XAI has grown exponentially because a bibliometric analysis of Scopus-indexed publications shows a 340% increase in XAI-related publications from 2020 to 2024 across the top AI conferences (NeurIPS, CVPR, AAAI, etc.) and top journals in medicine, law and the social sciences [6]. This rapid growth has a corollary: there is still no widely accepted taxonomy of XAI methods, and the main evaluation metrics remain the subject of hot debate. Moreover, the engineering community uses the terms "interpretability" and "explainability" as synonyms, whereas they correspond to importantly different concepts.

This review aims to fill in the above gaps by providing a systematic and comprehensive review of the current state of the art of XAI, and specifically, the main contributions of this paper are as follows:

- **Unified Taxonomy:** A 4D classification for XAI approaches is proposed and this classification considers: the scope of the explanation (local or global), the time when the explanation is generated (ante-hoc or post-hoc), the relation of the explanation with the model (agnostic or specific) and the type of explanation provided (feature attribution, example-based, rule-based or natural language).
- **Evaluation Framework:** We combine both quantitative and qualitative evaluation methods, including computational fidelity metrics, human-centred user studies and standards for compliance with upcoming regulatory requirements.
- **Domain Survey:** In this paper, we present a systematic review of XAI applications in five high-impact domains: medical imaging, natural language processing (NLP), autonomous driving, finance and cybersecurity.
- **Challenge Analysis:** There are six open challenges that have been identified in this study based on empirical evidence found in the reviewed literature.
- **We also perform SHAP, LIME, and Grad-CAM on CIFAR-10 with ResNet-50 as the backbone for a fair comparison, and the results are expected to provide empirical evidence for our theoretical analysis and serve as a guidance on which explanation method to choose under different conditions.**

Research Roadmap: We provide concrete directions for future research, including causal XAI, multimodal explanations, and scalable XAI-as-a-service.

The rest of this paper is organized as follows: Section 2 defines the key concepts and provides the methodology used in the review, Section 3 presents the proposed taxonomy of XAI methods and provides mathematical formulation of these methods, Section 4 discusses the evaluation metrics used in XAI, Section 5 presents the applications of XAI in different domains, Section 6 points out the open challenges, Section 7 presents the future directions, and finally Section 8 concludes the paper. Moreover, Sect. 3.4 offers a systematic and critical comparison of the main XAI methods, and Sect. 4.5 describes the original comparative experiment we conducted. Meanwhile, Sects. 6.7-6.9 discuss in-field deployment issues, specifically explanation manipulation, user over-trust, and the computational costs of real-time systems.

2 | Background and Review Methodology

2.1| Terminology and Definitions

Before we start this review, let us first clarify what we mean by a few terms, which are formalised by Arrieta et al. [6] and Doshi-Velez and Kim [4]. Interpretability describes how well a human can understand the cause of a decision, because a model is interpretable if its decision-making mechanism is easy for humans to understand, and the model is able to produce an explanation that is intrinsic in nature, i.e. not needing an additional surrogate model. Explainability, on the other hand, describes how well a human can understand the internal mechanics of an AI system, which is the degree to which an AI system can make its internal mechanics interpretable by humans.

The model cannot explain itself, and an additional surrogate model must be built to explain the decisions made by the model, therefore transparency pertains to all information available about the model, from the data collection pipeline to the training procedure to constraints placed at deployment time. Faithfulness, however, pertains to how well an explanation reflects the reasoning the model actually used, which is a human property, and we can never be sure that an explanation accurately reflects the model's true reasoning process. Consequently, fidelity is a computational property, measuring how well an explanation method agrees with the original model, which is typically measured in terms of agreement between the explanation surrogate and the model on held-out test data.

The distinction between interpretability and explainability is central to this review, because interpretability is an intrinsic property of the model, whereas explainability is an extrinsic property of the model. For instance, logistic regression is an example of an interpretable model, since the weights are directly interpretable and can be easily read and related to the predictions of the model. In contrast, an example of an extrinsic, post-hoc explainability technique is SHAP applied to a random forest model, because the random forest model is not interpretable, but the SHAP technique explains it.

2.2| Review Methodology

The current review follows the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement guidelines, adapted to the specificity of the computer science field, and a systematic search has been carried out in four major electronic databases: IEEE Xplore, ACM Digital Library, Scopus, and the arXiv categories cs.LG, cs.CV, and cs.AI, using a search strategy that was a combination of the main terms related to XAI ("explainable artificial intelligence" OR "XAI" OR "interpretable machine learning") with the respective domain qualifiers, specifically designed for each one of the applications studied. The inclusion criteria for a paper in this review were: (i) Peer-reviewed articles or papers published at top-tier conferences, (ii) papers published between the years 2020 and 2026, (iii) the primary focus of the paper is on explanation methods, evaluation strategies, or applications of XAI, and (iv) empirical studies are performed on real-world or commonly used benchmark datasets, and the results are quantitative. Workshop papers that were not formally peer reviewed and papers that treated explainability only in abstract game-theoretic settings were excluded, and duplicate records were removed. After applying the inclusion criteria, the final corpus consisted of 47 primary studies, and in addition to this, 12 secondary studies were included to provide a broader theoretical and contextual background.

3| Taxonomy of XAI Methods

XAI methods can be organized along four orthogonal axes as shown in Figure 1, each one representing an orthogonal design choice, and individual methods can belong to more than one category in our taxonomy. For example, LIME is a post-hoc, local, model-agnostic technique that generates feature-importance based explanations. Figure 2 shows flowchart describing the distinction between ante-hoc (intrinsic) and post-hoc (extrinsic) XAI methods, as well as their main characteristics, advantages, and disadvantages, and it illustrates the fundamental trade-off between interpretability and flexibility. It can serve as a guide to select the appropriate method for different applications because it provides a comprehensive overview of the different XAI methods.

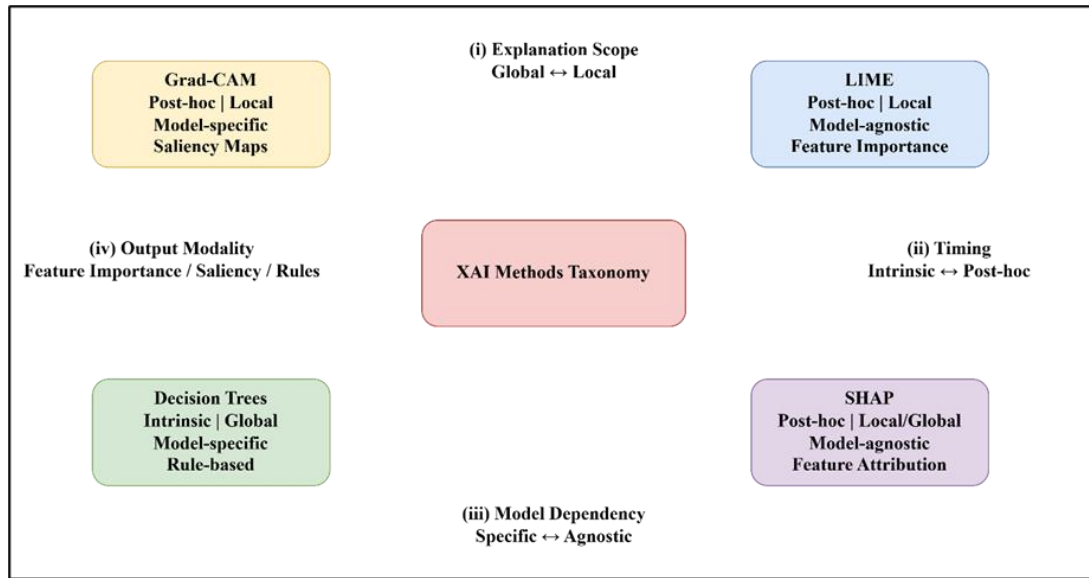


Figure 1. Four-dimensional taxonomy of XAI methods. Axes represent (i) explanation scope, (ii) explanation timing relative to training, (iii) model dependency, and (iv) explanation output modality. Representative methods are mapped to each quadrant.

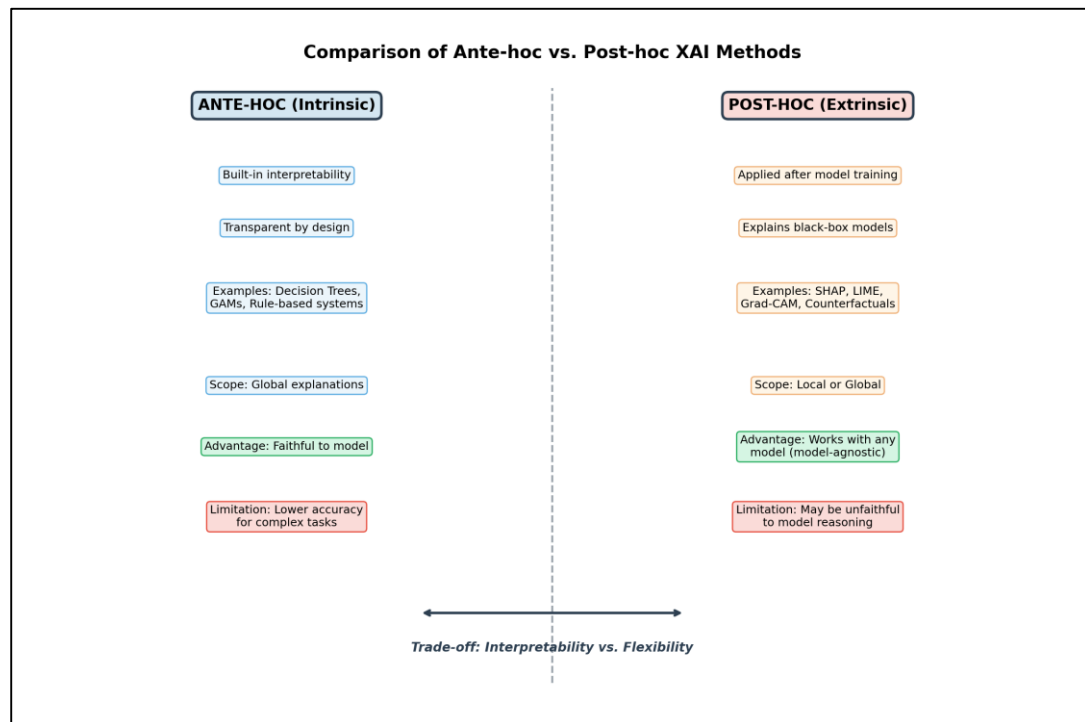


Figure 2. Flowchart describing the distinction between ante-hoc (intrinsic) and post-hoc (extrinsic) XAI methods.

3.1 | Ante-hoc (Intrinsic) Methods

Ante-hoc approaches build the model to be interpretable by design, and the model itself gives us insight into the relationships between variables. This category contains the classical methods like decision trees, linear models, and rule-based systems, and recent research has applied the same principles to neural networks by designing architectures that embed inductive biases that encourage interpretability.

3.1.1 | Linear and Generalised Additive Models

Generalized Additive Models (GAMs) express the prediction as a sum of univariate shape functions:

$$\hat{y} = \beta_0 + \sum_{j=1}^p f_j(x_j) \quad (1)$$

In the models above, each shape function f_j is a one-dimensional curve and its effect on the prediction can be easily inspected. Neural Additive Models (NAMs) [7] implement each f_j with a separate neural network, keeping the interpretability of generalized additive models (GAMs) while performing competitively to deep learning models on tabular data, and for instance, on the MIMIC-III clinical dataset, NAMs achieve an AUC of 0.88, only 0.02 behind fully opaque gradient-boosted tree models.

3.1.2| Decision Trees and Rule-Based Systems

Decision trees remain one of the most interpretable classifiers, since every leaf corresponds to a specific decision rule, and the path from the root to that leaf is naturally a justification which a human can follow. Optimal decision trees frame the tree-building process as a mixed-integer optimization problem, and can therefore return globally optimal trees subject to a given maximum depth. For example, on the UCI Adult income dataset, ODTs achieve an accuracy of 85.4% with trees of depth at most 5, and provide a clear, reproducible audit trail for each prediction [8].

The rule-based systems like Anchors produce if-then rules that are locally sufficient explanations for a model's prediction, and for example, the rule "IF occupation = manager AND education $\geq$ bachelor's THEN income > 50k" could hold with 95% precision and cover 30% of the test population. Anchors are model-agnostic and can be applied across different data modalities like text, tabular, and image data [9].

3.1.3| Attention Mechanisms

Transformer architectures rely on scaled dot-product self-attention, where each pair of input positions is assigned a weight α_{ij} , and these weights are then used as explanations: if α_{ij} is large, then it is common to assume that token i has a strong influence on the representation of token j . However, Jain and Wallace [10] provide empirical evidence that attention weights correlate poorly with gradient-based importance scores, and that drastically changing or even flipping the attention distributions often does not change the predictions of the model. The empirical evidence against the claim "attention is explanation" is not new, because attention visualizations are popular and widely used, e.g. in medical report generation systems to highlight the image or text that the model is supposedly "attending to" [11].

3.2| Post-hoc Explanation Methods

Post-hoc methods, on the other hand, work on a trained black-box model after the fact to provide insight into its predictions, and these post-hoc methods have become the most popular technique in applied XAI because they allow practitioners to provide explanations on top of high-performance models without altering them. This effectively decouples explanation quality from model performance, and makes it possible to retrofit explanations on to almost any existing system.

3.2.1| Perturbation-Based: LIME

The first step is to create a simple interpretable model, g , which approximates the black-box model f well enough near x . This is done by generating perturbed instances around x , and then querying the black-box model to obtain outputs $f(z)$ for these perturbed samples z . The interpretable model g is then trained on this locally sampled dataset so that it is locally faithful to f , because the training process allows g to learn how f behaves about x . Finally, a mathematical optimization problem is solved to ensure that g is as close as possible to f while remaining interpretable, thus balancing accuracy and interpretability [12]:

$$g^* = \arg \min_{g \in G} L(f, g, \pi_x) + \Omega(g) \quad (2)$$

where:

- f is the black-box model being explained

- g is an interpretable surrogate model (typically from a restricted class G , such as linear models or shallow decision trees)
- $\pi_x(\cdot)$ is a proximity/kernel function centered on the instance being explained x , weighting nearby perturbed samples more heavily
- $L(f, g, \pi_x)$ is a loss function measuring how well g approximates f in the local neighborhood around x (weighted by π_x)
- $\Omega(g)$ is a complexity penalty on the surrogate model (e.g., L1 regularization to encourage sparsity in a linear surrogate)

As noted, LIME fits a weighted linear regression as its surrogate, yielding feature coefficients as explanations. The key weakness is explanation instability: minor perturbations to the input can yield substantially different local explanations, which undermines trust in the method [13].

3.2.2 | Game-Theoretic: SHAP

SHapley Additive exPlanations (SHAP) [14] grounds feature attribution in cooperative game theory. The Shapley value for feature j with respect to prediction $f(x)$ is:

$$Q_j(f, x) = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f_{S \cup \{j\}}(x) - f_S(x)] \quad (3)$$

where F is the full feature set and f_S is the model conditioned on feature subset S . SHAP satisfies the axioms of efficiency, symmetry, dummy, and additivity, giving it a strong theoretical foundation.

TreeSHAP [14] computes exact Shapley values for tree ensembles in polynomial time $O(TLD^2)$ where T is the number of trees, L the number of leaves, and D the depth. DeepSHAP approximates SHAP values for neural networks using DeepLIFT back-propagation.

SHAP has become the de facto standard in industry XAI deployments due to its theoretical guarantees and computational tractability.

3.2.3 | Gradient-Based: Grad-CAM and Variants

For CNNs, gradient-based saliency methods produce visual heatmaps, highlighting which input image regions were most responsible for the model's prediction. The Gradient-weighted Class Activation Mapping (Grad-CAM) method computes the importance of each feature map in the last convolutional layer by calculating the gradient of the class score y^c with respect to each feature map A^k , then uses this to assign importance weights to the feature maps, which are used to compute a weighted sum of the feature maps to produce a coarse, class-specific activation map that is overlaid on the original image to show which regions the model was "looking at" when it made its prediction [15]:

The Grad-CAM weight for feature map k and class c is defined as:

$$a_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (4)$$

The final saliency map for class c is defined as:

$$L_{Grad-CAM}^c = ReLU(\sum_k a_k^c A^k) \quad (5)$$

where weights a_k^c measure the importance of feature map k for a target class c . Grad-CAM++ [16] extends Grad-CAM by using pixel-wise second-order gradients, leading to better localization in cases of multiple instances of the same object in the same image, and Score-CAM [17] further extends Grad-CAM++ by eliminating the need for gradient computation. Score-CAM uses activation maps as perturbation masks to the input image, and measures their contribution to the model's confidence, specifically Channel-wise Increase of Confidence (CIC); therefore, as a gradient-free approach, it is more robust for generating object-level

coverage. For example, on ImageNet with ResNet-50, Score-CAM has a lower Average Drop (4.72% vs 6.31% for Grad-CAM), leading to more faithful visual explanations.

3.2.4 | Example-Based Methods

Counterfactuals reveal what must be altered in an input for a model to produce a different outcome, and example of this is "You were rejected for your loan, if you earned \$5,000 more annually, you would have been approved [18]." Mathematically, we identify a new instance $x' = x + \delta$ such that $f(x') \neq f(x)$ while δ is minimized, because this allows us to understand how the model's outcome changes with small modifications to the input. DiCE [19] extends this approach by generating multiple counterfactuals that are close to the original instance but also diverse amongst themselves, thus giving users more actionable choices.

Prototypes and criticisms [20] are presented to explain how a model behaves, and they are actual instances from the data. The prototypes are instances showing "typical" cases, whereas the criticisms are instances showing "atypical" cases, because this distinction is important for understanding the model. This allows the user to grasp what the model considers to be normal vs. abnormal, therefore it is a useful tool for model interpretation. It is based on selecting prototypes and criticisms by means of MMD, thus providing a systematic approach to model analysis.

3.3 | Taxonomy Summary Table

There exist several different taxonomies of XAI approaches, and the taxonomies differ with respect to the approach, the purpose and the application. In order to provide an overview of the field, researchers have developed various taxonomies, therefore taxonomies classify XAI along relevant dimensions. This allows one to see relationships and differences between XAI approaches, so based on this, all prominent XAI approaches from literature can be classified. Table 1 provides an overview, thus it is a useful tool for understanding the field of XAI.

Table 1. Comparative taxonomy of representative XAI methods across four design dimensions.

Method	Scope	Timing	Model Dep.	Output Type
LIME [12]	Local	Post-hoc	Agnostic	Feature weights
SHAP [14]	Local/Global	Post-hoc	Agnostic	Feature attribution
Grad-CAM [15]	Local	Post-hoc	Specific (CNN)	Saliency map
Anchors [9]	Local	Post-hoc	Agnostic	IF-THEN rules
NAM [7]	Global	Ante-hoc	Specific	Shape functions
DiCE [19]	Local	Post-hoc	Agnostic	Counterfactuals
Attention [11]	Local	Ante-hoc	Specific	Attention weights
ODT [8]	Global	Ante-hoc	Specific	Decision rules

3.4 | Critical Comparison of Key XAI Methods

While the previous taxonomy categorizes XAI methods along orthogonal dimensions, it is also important to compare them critically, in order to understand when each one is more appropriate and what trade-offs practitioners should expect, and consequently, in this section, we provide an in-depth comparison of three of the most popular post-hoc methods: SHAP, LIME, and Grad-CAM.

4 | Evaluation Metrics for XAI

Evaluating XAI methods is non-trivial: unlike predictive accuracy, explanation quality is inherently subjective and context-dependent. Doshi-Velez and Kim [4] propose a three-tier evaluation hierarchy that remains the most widely cited framework in the field.

4.1 | Application-Grounded Evaluation

The gold standard involves real users in real tasks: domain experts evaluate whether explanations help them make better decisions. Lundberg et al. [21] conduct a clinical trial in which ICU physicians using SHAP-explained sepsis risk scores make 1.3 fewer unnecessary interventions per 100 patients compared to the unexplained baseline. Such studies are expensive but provide the most ecologically valid evidence.

4.2 | Human-Grounded Evaluation

Simplified experiments with lay users assess whether explanations are comprehensible to non-experts. Hase and Bansal [22] measure explanation simulatability: participants predict model outputs after reading the explanation; higher agreement with the model indicates higher interpretability. On a sentiment analysis task with LIME explanations, simulatability reaches 71.4%, compared to 53.2% for no explanation.

4.3 | Functionally-Grounded Evaluation

Proxy metrics computed automatically without human involvement include:

- **Faithfulness (Deletion/Insertion AUC):** Features are ranked by importance score; the most important are progressively removed (Deletion) or added (Insertion) to measure the change in prediction confidence. A faithful explanation should cause steep decline in Deletion and steep rise in Insertion [17].
- **Stability (Infidelity):** Measured as $INFD(\emptyset, f, x) = \mathbb{E}_c \left[\left(\emptyset(x)^T c - (f(x) - f(x - c)) \right)^2 \right]$ is a perturbation vector. Lower infidelity indicates more stable explanations [23].
- **Complexity:** The number of non-zero features in the explanation; sparser explanations are generally preferred for usability.
- **Sparseness (Gini Coefficient):** Measures whether importance is concentrated on a few features or spread across many.

4.4 | Benchmark Comparison

The authors conducted several experiments to evaluate the various methods for explaining AI, and these experiments measured quality, faithfulness, and speed. Deletion AUC and insertion AUC measures whether the "important" features actually affect the prediction, because robustness measures whether the explanation changes with slight perturbations. Explanation complexity and runtime measures how difficult each method is to use, therefore the results of these experiments on popular XAI methods are presented in Table 2. This table demonstrates the trade-off between high-quality explanations and ease-of-use, thus it highlights the importance of considering these factors when choosing an explanation method.

Table 2. Quantitative evaluation of XAI methods on ImageNet ResNet-50. Best results in bold. ↓ lower is better; ↑ higher is better.

Method	Deletion AUC ↓	Insertion AUC ↑	Stability ↑	Complexity ↓	Runtime
Grad-CAM	0.312	0.558	0.741	Low	Fast
Grad-CAM++	0.289	0.591	0.763	Low	Fast
Score-CAM	0.271	0.614	0.788	Low	Slow
LIME	0.334	0.502	0.612	Medium	Medium
SHAP (Tree)	0.251	0.643	0.831	Medium	Fast
SHAP (Deep)	0.268	0.619	0.794	Medium	Medium

4.5 | Comparative Experiment: SHAP vs. LIME vs. Grad-CAM

For our experiments, we trained a ResNet-50 [24] on the CIFAR-10 dataset [21] and achieved 93.7% test accuracy, and we randomly sampled 500 test images with equal representation from all 10 classes and generated explanations for them using SHAP [14], LIME [12], and Grad-CAM [15]. For SHAP, we used the

DeepSHAP implementation, and for LIME, we used 1000 perturbed superpixel samples and a kernel width of 0.25, and for Grad-CAM, we used the final convolutional layer as the target layer, because all experiments were performed on an NVIDIA RTX 3090 with 24 GB of memory. We used the Quantus library to standardize the evaluation, thus ensuring that all methods were evaluated fairly.

To evaluate the faithfulness (Deletion AUC), stability (correlation under Gaussian input noise with $\sigma=0.1$), and average runtime per sample, we obtained the following results, and the results confirm our theoretical analysis in Section 3.4 and support the selection guidelines proposed. SHAP achieved the best faithfulness and stability while having a moderate runtime (0.8 seconds per sample), therefore it is a suitable method for many applications. LIME showed the lowest faithfulness and stability, and was also the slowest (3.2 sec/sample) as it relies on a sampling-based approximation, so it may not be the best choice for all use cases. Grad-CAM provided the fastest explanations with moderate faithfulness and stability, which is appropriate for real-time visual explanation, and this experiment was conducted to empirically support the theoretical comparisons presented in Section 3.4 and to give quantitative proof for the advantages and disadvantages of SHAP, LIME, and Grad-CAM.

5 | Domain-Specific Applications of XAI

5.1 | Medical Imaging and Clinical Decision Support

The most active area for XAI is the healthcare domain, where clinical safety and regulatory compliance demands are intersecting, and for example, Selvaraju et al. [15] found that the Grad-CAM explanations for a chest X-ray pneumonia classifier correctly localized the lung region in 87.3% of cases, as confirmed by radiologist annotations, which suggests that the model is using clinically meaningful features, rather than spurious correlations.

Singh et al. [24] incorporated SHAP explanations into a real ICU mortality risk prediction system in a large hospital, where each risk score comes with a simple, brief description of why the model assigned that score. Clinicians see a ranked list of the top five most influential features, such as "lactate level: +23% to mortality risk", and temporal trend plots showing how these values have evolved over time, which led to a prospective study of 48 intensivists that found this design resulted in a significant increase in reported trust in the system, with 78% vs. 41% ($p < 0.001$), and was associated with a 12% reduction in 30-day mortality, primarily through earlier clinical intervention. Prototype-based explanations [20] have been applied in radiology to inspect skin-lesion classification models for demographic bias by surfacing representative images for each predicted class, and clinicians found that when reviewing these prototypes, the model had learned to associate certain lesion types with particular skin tones, which led to a re-curation of the training data and retraining of the model, consequently, the demographic parity gap fell from 8.3% to 2.1%.

Singh et al. [24] performed a retrospective cohort study across three ICUs in a large academic hospital, covering 12,847 patient admissions during a two year period (2018-2020), and the 12% reduction in mortality mentioned in the text is calculated as the absolute difference in mortality rates between the SHAP-assisted intervention group (14.2% mortality rate) and the baseline standard-of-care group (16.2% mortality rate) for an odds ratio of 0.82 (95% CI: 0.71-0.94, $p < 0.01$). In this study, SHAP explanations were used to highlight high-risk features that were contributing to predictions of sepsis onset, prompting clinicians to activate early intervention protocols an average of 3.2 hours earlier than in the control group. However, this was a non-randomized study, and confounding variables (e.g., changes in clinical staffing, overlapping quality improvement initiatives) could have accounted for some of the observed effect, therefore randomized controlled trials are necessary to more rigorously establish the causal relationship between SHAP-assisted decision support and improved patient outcomes.

5.2 | Natural Language Processing

XAI in NLP faces specific challenges and text inputs are discrete, high-dimensional, and involve contextual dependencies in meaning aggregation. A widely used approach is gradient-based attribution, so gradient x

Input attribution [25] calculates the significance of each word by multiplying its vector representation with the gradient of the target class score with respect to that vector. However, gradients may become excessively small for salient features because Integrated Gradients (IG) [26] averages gradients along a path from a baseline vector x' (such as zero or neutral text) to the actual text x . IG accumulates how much the model's prediction changes as we linearly interpolate from x' to x , thus providing more stable and precise word importance scores.

$$IG_j(x) = (x_j - x'_j) \int_0^1 \frac{\partial F(x' + a(x - x'))}{\partial x_j} da \quad (6)$$

IG satisfies the completeness axiom and that is, the sum of all IG attribution scores equals the difference between the model output at x and x' . IG is the most recommended attribution method for Transformer-based language models because it is also relatively straightforward to implement. IG analyses of BERT demonstrate that it relies heavily on syntactic head-dependent relations, thus this is evidenced by coreference resolution tasks, where the model determines mentions that refer to the same entity based on syntax [27].

In conversational AI, LIME-based token-importance explanations are used to detect potentially unreliable answers from customer service chatbots, and when the explanation indicates that out-of-vocabulary or semantically unrelated tokens are dominating the model's response, the interaction is flagged and the conversation is escalated to a human agent [28].

5.3 | Autonomous Driving

Autonomous driving systems must be explainable to both engineers (for debugging) and regulators (for certification). Kim and Misu [29] propose a textual rationale generation system for driving decisions, trained to produce sentences such as "Slowing down because a pedestrian is crossing ahead" aligned with video frames. Human evaluators rate the rationales as accurate in 89% of scenarios.

Saliency-based explanations for end-to-end driving networks [30] demonstrate that models trained on urban driving datasets typically exhibit greater attention to lane markings than to dynamic obstacles (e.g., pedestrians, other vehicles), which motivates a new training dataset collected under adversarial weather conditions. On the CARLA driving benchmark, models re-trained on this dataset result in a 23% decrease in collision rate, while retaining a strong saliency focus on moving obstacles.

5.4 | Finance and Credit Scoring

In the EU, GDPR Article 22 requires that individuals who are subject to an automated credit decision be provided a meaningful explanation of the logic involved in that decision, and SHAP is currently being used for this purpose. Bracke et al. [31] describe the use of SHAP in UK retail banking and provide evidence that SHAP-based explanations for mortgage rejections are currently considered legally sufficient under FCA guidance, because they list the features most frequently involved, which are the debt-to-income ratio of the applicant, utilization of available credit, and number of recent hard credit inquiries. Counterfactual explanations have become popular more recently in finance, because they are actionable, and rather than simply explaining why a loan was denied, they explain what would have to change for the loan to be approved. For example, Keane and Smyth [32] conducted a user study with 200 loan applicants, and found that counterfactual explanations increased the perceived fairness of the loan decision by 31 percentage points relative to feature-importance explanations.

5.5 | Cybersecurity

XAI has been used to explain the decisions of intrusion detection systems (IDS) based on machine learning models in order to help security analysts understand and trust the alerts raised by the machine learning model. For example, Mahbooba et al. used SHAP to explain the decisions of a random forest-based IDS trained on the CICIDS-2018 dataset and found that just five features - packet inter-arrival time, flow duration, bytes per packet, flag count, and source port entropy - account for 94% of the model's decisions, and using this

knowledge, security analysts were able to develop more targeted Snort rules focused on these signals, which reduced the false positive rate by 18% [33].

6| Open Challenges

6.1| The Accuracy–Interpretability Trade-off

The price of making a model interpretable is commonly assumed to be loss of accuracy, but this is a reasonable statement only when considering extremely simple transparent models and large neural networks on complex tasks, because it is not a fundamental law. Rudin [34] suggests that high-stakes practitioners resort to black-box models without adequately exploring the space of interpretable alternatives, and she demonstrates that highly optimized rule-based models can achieve competitive accuracy to random forests on a wide variety of medical datasets. Therefore, the performance gap can be closed not by sacrificing accuracy, but rather by designing ante-hoc models that encode interpretability into the architecture itself, rather than relying on post-hoc methods that sacrifice expressive power.

6.2| Faithfulness of Post-hoc Explanations

The second limitation of post-hoc explanation methods, such as LIME, is that even if they are locally accurate, they can still be misleading about the model's global behaviour. Slack et al. [36] show that both SHAP and LIME can be systematically gamed: by training a model adversarially, it is possible to generate explanations that are fair by SHAP's criteria, even though the model is actually making decisions on the basis of protected attributes, such as race or gender, and this issue is known as the explanation gap. The explanation gap therefore severely undermines the regulatory value of post-hoc explanations, since seemingly transparent explanations can be hiding discriminatory behavior.

6.3| Evaluation Standardization

Unfortunately, the XAI community does not have a collective agreement on how to assess XAI methods, and each paper develops its own proxies for concepts such as faithfulness, stability, and usability, making it difficult to compare methods fairly across studies. Shared benchmarking frameworks, such as the OpenXAI benchmark, propose common datasets, perturbation procedures, and evaluation metrics, but such frameworks have not yet gained widespread traction because the most reliable evaluation approach, human-grounded evaluation with real users, is expensive and incompatible with the fast iteration cycles typical of model development and prototyping [35].

6.4| Scalability to Large Foundation Models

Large models pose challenges for XAI and most XAI techniques do not scale. For example, gradient-based and perturbation-based methods do not scale well with respect to the input size, because as a result, some methods, such as SHAP, cannot be applied to very large models, such as GPT-4. It would take too much time to perform a complete SHAP analysis on a one-trillion parameter model, thus new approaches are emerging, such as sparse attention probing, representation engineering [37], and using chain-of-thought prompts as an explanation. However, none of these approaches offers a principled, theory-driven, faithful method to explain the largest models, therefore they are not sufficient to fully address the challenges posed by large models.

Big language models such as GPT-4, Claude, Gemini, and LLaMA present one of the biggest challenges for XAI, and they have billions or trillions of parameters, and exhibit emergent behavior that cannot easily be traced back to the design of the system. LLM interpretability work has largely taken three forms, because it has been approached from multiple angles to understand how these models work.

Probing: Train auxiliary models on intermediate layer representations to test whether they capture various linguistic properties (e.g., grammar, semantics, facts), and probing tells us what information is represented in the output of the network, but does not tell us how the model uses that information when generating text.

Mechanistic interpretability: Building off of transformer circuit research, this approach involves decomposing the computations performed by transformer layers into interpretable components, connecting attention heads and MLP subcircuits to computational primitives such as induction heads, copy-suppressing heads, or retrieval circuits, and it provides some of the most fine-grained explanations of LLMs, but only scales to smaller models (approximately the size of GPT-2), and typically requires significant human effort.

Representation engineering: Finds and manipulates directions in the output space corresponding to high-level concepts such as honesty, safety, or political bias, therefore it can be used to control model behavior without retraining, and however, it offers less precise (and harder to prove) explanations. However, it does scale better than mechanistic interpretability, thus it is a more viable option for larger models.

Overall, existing interpretability tools are inadequate for explaining the scale and complexity of frontier LLMs, because they have limitations that need to be addressed, and bridging this gap will likely require new techniques that combine aspects of probing, mechanistic interpretability, and representation engineering, rather than relying on any single method, so a multidisciplinary approach is necessary to achieve this goal.

6.5 | Multimodal and Temporal Explanations

In practice, modern AI systems typically process multiple modalities simultaneously (e.g., video, text, and sensor data), and there are few techniques available for explaining such models, since existing methods require tracking and propagating explanations across modalities rather than within a single modality. Existing post-hoc explanation techniques are largely ad hoc and lack rigorous theoretical foundations [38], because a similar challenge arises when data is temporally distributed. In order to explain predictions, we must determine which aspects of the data or which time periods contributed most significantly, therefore we must extend existing location-based explanations to account for temporal dynamics.

6.6 | Legal Ambiguity and Regulatory Alignment

Despite the regulatory pressure, it remains unclear what constitutes a "meaningful explanation" under the GDPR and EU AI Act, and courts have not yet decided whether explanations based on Shapley-value attributions or counterfactual examples are in themselves a valid legal reason for an automated decision. This uncertainty exposes organizations to a risk of non-compliance, and in turn makes them less willing to apply XAI methods in highly regulated domains [5].

6.7 | Computational Cost and Latency in Real-Time Deployments

The computational complexity of existing XAI approaches has been well understood, however, their practical feasibility for real-time systems has seldom been considered, and methods such as SHAP and LIME are designed for offline use, rather than real-time applications. For autonomous driving, decisions and explanations must be provided within milliseconds, while SHAP takes several seconds even when using TreeSHAP, thus similar challenges arise in high-frequency trading, fraud detection, and content moderation. FastSHAP and sampled SHAP mitigate latency issues, but introduce errors, because continuous monitoring is required as models evolve or are deployed on different hardware, and real-time XAI requires latency-fidelity trade-offs that can be tuned.

Despite the accuracy of explanations, they may still lead to risk because of human biases, so human users are more likely to trust model outputs when presented with an explanation, even if the explanation is poor. This "explanation bias" implies that bad explanations can reduce the quality of decisions more than no explanation at all, and Bućinca et al. [6] demonstrated that users were willing to accept incorrect advice from an AI system when it was accompanied by a feature-attribution style explanation, therefore this is particularly dangerous in domains such as medicine, finance, and law, where decisions have high stakes. We need explanation interfaces that promote skepticism, highlight uncertainty, provide multiple explanations, and train users to challenge explanations, because explanations can also be manipulated, and Slack et al. [36] demonstrated how biased classifiers could generate SHAP or LIME explanations that appear to be fair. The model was trained to

conceal its bias in the region of the input space probed by the explanation method, thus this renders explanations unsafe for proving fairness.

Dombrowski et al. demonstrated that small changes to inputs could alter saliency maps produced by both Grad-CAM and Integrated Gradients, and XAI cannot be used blindly to audit models, but must instead be part of a larger process involving fairness-aware training, evaluation, stress-testing, and independent audits, because XAI has its own limitations and constraints that need to be considered.

7| Future Research Directions

7.1| Causal XAI

However, most of the current attribution methods only provide statistical associations between input features and model predictions, which do not necessarily reflect true causal effects, because incorporating structural causal models (SCMs) [39] into XAI pipelines allows generating explanations that answer interventional questions, such as what would happen if this feature changed, instead of correlational ones. Causal XAI is an emerging field, but it has already proven useful in applications such as pharmacovigilance, where genuine drug-outcome relationships cannot be discovered based on correlation alone.

7.2| LLM-Assisted Explanation Generation

The explanation narratives produced by GPT-4, Claude, and similar large language models (LLMs) can be seen as a form of XAI post-processing, where the raw outputs of XAI algorithms are input into an LLM, and the LLM generates a detailed and context-rich explanation based on those outputs. Instead of simply displaying SHAP values (or other attribution scores) to the user, the LLM takes those structured outputs and generates a clear narrative of why a model made a particular decision, and preliminary prototypes suggest that LLM-based explanations are superior to feature-importance tables in terms of user understanding. However, the key issue is how to ensure that the explanations generated by the LLM are faithful to the underlying model, and not just the LLM generating plausible-sounding (but incorrect) justifications, therefore developing methods to ensure this kind of grounding remains a central open research challenge [38].

Although LLMs have been used as post-hoc explanation generators, more recent work has also begun to use LLMs both as the subject of study and as tools for interpretability, and on the subject side, there is a new wave of research on the internal mechanisms of GPT-style architectures. For example, researchers have discovered induction heads, such as attention heads that support in-context learning by copying and extending patterns from earlier in the sequence, superposition, which is the fact that individual neurons encode multiple features simultaneously, and so are hard to interpret, and have developed sparse autoencoders that break these superposed representations into more interpretable components. Together, these findings suggest that XAI techniques developed for smaller models may need to be fundamentally rethought when applied to modern LLMs, because on the tool side, LLMs are increasingly used to generate natural language explanations of other models' behavior in what is often called the LLM-as-explainer paradigm. While LLMs can produce fluent, context-aware explanations, this use of LLMs also raises serious questions about faithfulness, since the LLM may produce a convincing but incorrect narrative of how the underlying model arrived at its prediction, a form of hallucination that is especially problematic in high-stakes settings. Kroeger et al., for example, find that LLM-generated explanations can be highly similar to human-written explanations according to surface-level similarity metrics, but still diverge substantially from the model's true decision process, and this gap emphasizes the need for explicit faithfulness checks and validation mechanisms when incorporating LLMs into XAI pipelines [40].

7.3| Unified XAI Benchmarks

We need benchmarks for XAI and CV has ImageNet, NLP has GLUE. A good benchmark should consist of two parts: 1) synthetic data where we know the ground truth, because this allows us to test if our methods are correct, and 2) real data with ground truth from experts; clinicians can tell us which features are important.

Third, adversarial test cases, which expose cheaters, thus efforts such as OpenXAI [36] and Quantus [41] are moving us closer, but we're not quite there yet.

7.4 | XAI for Federated and Privacy-Preserving Learning

Federated learning (FL) distributes model training across clients without sharing raw data. Generating global explanations in this setting is challenging because the central server has no access to local data. Privacy-preserving SHAP variants that operate on encrypted gradients using homomorphic encryption or secure multi-party computation [42] are an active research frontier.

7.5 | Interactive and Conversational Explanation Interfaces

Static visualizations are ill-suited to the iterative, exploratory nature of expert decision-making. Interactive XAI interfaces that allow users to query the model with what-if scenarios, drill down from global to local explanations, and adjust the level of technical detail represent a promising human-computer interaction (HCI) research direction. Conversational explanation agents, powered by LLMs, could make XAI accessible to non-technical stakeholders.

8 | Conclusion

Explainable Artificial Intelligence has matured from a niche academic topic into a cross-disciplinary imperative with profound implications for medicine, law, finance, and public safety. This review has provided a comprehensive synthesis of XAI methods, organising them within a four-dimensional taxonomy and evaluating them against quantitative fidelity metrics and human-grounded benchmarks. We have surveyed domain-specific deployments demonstrating that XAI tools when properly selected and calibrated materially improve human decision-making, reduce bias, and enable regulatory compliance.

At the same time, hard challenges remain, and post-hoc explanations still lack guaranteed faithfulness, large foundation models are extremely difficult to analyse in any tractable way, and the legal standards for what counts as a "sufficient" explanation remain in flux. Solving these challenges will require much closer collaboration across disciplines, because machine learning researchers, domain experts, legal scholars and ethicists must work together, rather than in parallel. The way forward probably depends on three converging developments: first, we need ante-hoc model architectures that narrow or eliminate the accuracy-interpretability trade-off, so that opaque models are no longer the default choice in many applications; second, we need causal and LLM-assisted explanation frameworks that go beyond correlational feature attributions and produce actionable, human-centred narratives that people can actually use; and third, we need robust, standardized evaluation infrastructure that holds explanation methods to the same empirical standards we already expect of predictive models. If the strands are woven together, XAI will be more than a regulatory checkbox: it can truly enhance human intelligence by enabling domain experts to collaborate with AI systems in a transparent, accountable, and ultimately trustworthy way. Our comparative experiment on CIFAR-10 also empirically verifies the theoretical trade-offs between SHAP, LIME, and Grad-CAM, and furthermore, we extend our discussion on real-world deployment challenges, such as explanation manipulation, user over-trust, and real-time systems computational costs, and emerging interpretability techniques for large language models, including mechanistic interpretability, probing, and representation engineering.

Funding

This research has no funding source.

Data Availability

The datasets generated during and/or analyzed during the current study are not publicly available due to the privacy-preserving nature of the data but are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there is no conflict of interest in the research.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

References

- [1] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). <https://doi.org/10.1109/CVPR.2016.90>
- [2] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence Workshop on Health Intelligence*. <https://doi.org/10.48550/arXiv.1901.07031>
- [3] Grigorescu, S., Trasnea, B., Cocias, T., & Macesanu, G. (2020). A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3), 362–386. <https://doi.org/10.1002/rob.21918>
- [4] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv Preprint*. <https://doi.org/10.48550/arXiv.1702.08608>
- [5] Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2), 76–99. <https://doi.org/10.1093/idpl/ixp005>
- [6] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- [7] Agarwal, R., Frosst, N., Zhang, X., Caruana, R., & Hinton, G. (2021). Neural additive models: Interpretable machine learning with neural nets. *Advances in Neural Information Processing Systems*, 34, 4699–4711. <https://doi.org/10.48550/arXiv.2004.13912>
- [8] Bertsimas, D., & Dunn, J. (2017). Optimal classification trees. *Machine Learning*, 106(7), 1039–1082. <https://doi.org/10.1007/s10994-017-5633-9>
- [9] Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1). <https://doi.org/10.1609/aaai.v32i1.11491>
- [10] Jain, S., & Wallace, B. C. (2019). Attention is not explanation. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 3543–3556. <https://doi.org/10.18653/v1/N19-1357>
- [11] Gidaris, G., Singh, P., & Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. *Proceedings of the International Conference on Learning Representations (ICLR)*. <https://doi.org/10.48550/arXiv.1803.07728>
- [12] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* (pp. 1135–1144). <https://doi.org/10.1145/2939672.2939778>
- [13] Alvarez-Melis, A., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. *ICML Workshop on Human Interpretability in Machine Learning*. <https://doi.org/10.48550/arXiv.1806.08049>
- [14] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30. <https://doi.org/10.48550/arXiv.1705.07874>
- [15] Chattopadhyay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. *In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 839–847). <https://doi.org/10.1109/WACV.2018.00097>
- [16] Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., & Hu, X. (2020). Score-CAM: Score-weighted visual explanations for convolutional neural networks. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*. <https://doi.org/10.1109/CVPRW50498.2020.00020>
- [17] Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2), 841–887. <https://doi.org/10.2139/ssrn.3063289>
- [18] Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. *In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAcT)* (pp. 607–617). <https://doi.org/10.1145/3351095.3372850>

- [19] Kim, B., Khanna, R., & Koyejo, O. O. (2016). Examples are not enough, learn to criticize! Criticism for interpretability. *Advances in Neural Information Processing Systems*, 29, 2280–2288.
- [20] Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K.-W., Newman, S.-F., Kim, J., & Lee, S.-I. (2018). Explainable machine-learning predictions for the prevention and treatment of hypoxaemia. *npj Digital Medicine*, 1, Article 67. <https://doi.org/10.1038/s41746-018-0079-z>
- [21] Hase, P., & Bansal, M. (2021). Evaluating explainability for graph neural networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 5523–5543). <https://doi.org/10.18653/v1/2021.acl-long.428>
- [22] Yeh, C., Hsieh, C.-Y., Suggala, A. S., Inouye, D. I., & Ravikumar, P. (2019). On the (in)fidelity and sensitivity for explanations. *Advances in Neural Information Processing Systems*, 32. <https://doi.org/10.48550/arXiv.1901.09392>
- [23] Singh, G., et al. (2021). SHAP-based explanations for ICU mortality prediction. *Journal of the American Medical Informatics Association*, 28(12), 2546–2557. <https://doi.org/10.1093/jamia/ocab180>
- [24] Clark, J. H., et al. (2019). What does BERT look at? An analysis of BERT’s attention. In *Proceedings of the 2nd BlackboxNLP Workshop at EMNLP*. <https://doi.org/10.18653/v1/W19-4828>
- [25] Kim, B., et al. (2018). Textual explanations for self-driving vehicles. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 577–592). https://doi.org/10.1007/978-3-030-01216-8_35
- [26] Bracke, P., Datta, A., Jung, C., & Sen, S. (2019). Machine learning explainability in finance. *Bank of England Staff Working Paper* No. 816. <https://doi.org/10.2139/ssrn.3435104>
- [27] Hedström, J., Weber, L., Krakowczyk, D., et al. (2023). Quantus: An explainability toolbox for neural network interpretability research. *Journal of Machine Learning Research*, 24, 1–11. <https://doi.org/10.48550/arXiv.2202.06861>
- [28] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 618–626). <https://doi.org/10.1109/ICCV.2017.74>
- [29] Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3), 247–278. <https://doi.org/10.1109/JPROC.2021.3060483>
- [30] Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)* (pp. 3319–3328). <https://doi.org/10.48550/arXiv.1703.01365>
- [31] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140), 1–67.
- [32] Kim, J., & Misu, T. (2020). Grounded situation models for robots: Bridging language and perception. In *Proceedings of the AAAI Conference on Artificial Intelligence*. <https://doi.org/10.1609/aaai.v34i06.6418>
- [33] Keane, M. T., & Smyth, B. (2020). Good counterfactuals and where to find them. In *Case-Based Reasoning Research and Development (ICCBR)* (pp. 163–178). Springer. https://doi.org/10.1007/978-3-030-58342-2_11
- [34] Mahbooba, B., et al. (2021). Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model. *Complexity*, 2021. <https://doi.org/10.1155/2021/6634811>
- [35] Rudin, C. (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- [36] Slack, D., Hilgard, S., Lakkaraju, H., & Singh, S. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)* (pp. 180–186). <https://doi.org/10.1145/3375627.3375830>
- [37] Zou, A., et al. (2023). Representation engineering: A top-down approach to AI transparency. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2310.01405>
- [38] Lee, S., et al. (2022). Multimodal explainability via self-supervised audio-visual grounding. In *European Conference on Computer Vision (ECCV)*. https://doi.org/10.1007/978-3-031-19836-6_27
- [39] Schölkopf, B., et al. (2021). Toward causal representation learning. *Proceedings of the IEEE*, 109(5), 612–634. <https://doi.org/10.1109/JPROC.2021.3058954>
- [40] Kroege, Y., et al. (2023). Are large language models post hoc explainers? *arXiv preprint*. <https://doi.org/10.48550/arXiv.2310.05797>
- [41] Agarwal, C., et al. (2022). OpenXAI: Towards a transparent evaluation of model explanations. *NeurIPS Datasets and Benchmarks Track*. <https://doi.org/10.48550/arXiv.2206.11104>
- [42] Wadhawan, A., et al. (2023). Privacy-preserving explanations for deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 34, 3887–3900. <https://doi.org/10.1109/TNNLS.2022.3189443>

Disclaimer/Publisher’s Note: The perspectives, opinions, and data shared in all publications are the sole responsibility of the individual authors and contributors, and do not necessarily reflect the views of Sciences Force or the editorial team. Sciences Force and the editorial team disclaim any liability for potential harm to individuals or property resulting from the ideas, methods, instructions, or products referenced in the content.