International Journal of Sustainable Applied Science and Engineering (IJSASE)

Contents lists available at IJISASE

International Journal of Sustainable Applied Science and Engineering

journal homepage: IJISASE

Volume 2, No. 1, 2025

Beyond the Black Box: A Review of Quantitative Metrics for Neural Network Interpretability and Their Practical Implications

Maryam Esna-Ashari [a]

[a] *Department of the Property and Casualty Insurance, Insurance Research Center, Tehran, Iran*

ARTICLE INFO

Received: 2024/12/07

Revised: 2025/01/05

Accept: 2025/02/09

Keywords:

Neural network interpretability, Quantitative interpretability metrics, Fairness in AI, Fidelity, Explainability.

ABSTRACT

As neural networks continue to grow in complexity and find applications in critical domains such as healthcare, finance, and autonomous systems, the demand for transparent and trustworthy AI has never been greater. This paper provides a comprehensive review of quantitative metrics used to evaluate the interpretability of neural networks, focusing on key measures—fidelity, complexity, robustness, and sensitivity—and examining their respective advantages, limitations, and suitability across different model architectures. In addition, the review explores major challenges in interpretability assessment, including data quality, bias, scalability, and generalizability, while highlighting emerging approaches such as causal and interactive interpretability. By addressing these core issues and advancements, the paper aims to bridge the gap between high model performance and meaningful transparency, ultimately contributing to the development of more accountable and trustworthy AI-driven decision-making systems.

1. Introduction

Neural networks, particularly deep learning models, are increasingly being deployed in critical sectors such as healthcare, finance, and autonomous systems. In these high-stakes environments, the ability to understand and explain a model's decision-making process, commonly referred to as interpretability, has become imperative [4]. Interpretability not only enables stakeholders to trust and validate model outputs, but also ensures alignment with human values and ethical standards. For instance, in healthcare applications, model-driven decisions can carry profound consequences for patients, making transparent

[a] Corresponding author email address: esnaashari@irc.ac.ir (Maryam Esna-Ashari).

Available online 02/09/2025

3060-6349 /BGSA Ltd.

reasoning essential for both clinicians and regulatory bodies [11]. However, deep neural networks typically involve numerous laysers and vast numbers of parameters, which interact in complex and highly nonlinear ways. Tracing how specific inputs lead to specific outputs thus becomes a formidable challenge [33]. Unlike traditional machine learning algorithms such as decision trees or linear regressions, neural networks are often characterized as "black-boxes" due to the opacity of their internal processes. This opacity can mask spurious correlations or hidden biases that remain undetected, even if the model's predictions appear accurate on the surface [22]. Consequently, there is a growing need for robust interpretability metrics that can ascertain whether a model's internal logic aligns with human reasoning and domain knowledge. However, defining and developing such metrics is inherently difficult, as interpretability itself is a multifaceted and context-dependent concept [33]. Models optimized for maximum accuracy can sacrifice transparency, highlighting the inherent tension between performance and explainability. In response, researchers have proposed a variety of quantitative interpretability measures aimed at providing objective, repeatable, and scalable assessments. While qualitative methods—such as visualizations or narrative explanations—can offer valuable insights, they alone may not suffice in domains requiring rigorous validation over large datasets or ongoing monitoring [4, 56].

This paper aims to review and evaluate quantitative metrics that address these challenges by systematically examining their strengths, limitations, and applicability across a spectrum of neural network architectures and real-world applications. In addition to covering established metrics (e.g., fidelity, complexity, robustness), we consider advanced topics such as the dynamic, evolving nature of neural networks in online or streaming contexts. As models update their parameters in response to new data, their decision-making criteria can change over time, an aspect often overlooked by static interpretability approaches [31]. Furthermore, interpretability in ensemble models introduces another layer of complexity, where multiple networks or heterogeneous machine learning algorithms collectively influence the final decision. Hence, the development of innovative metrics capable of capturing both individual and collective model behavior has become increasingly crucial [54]. By highlighting these core challenges and exploring emerging methodologies, this article contributes to the ongoing effort to balance high performance with meaningful transparency in neural network–based systems. The ultimate goal is to increase accountability, foster user trust, and support responsible AI deployment in mission-critical domains.

1.1 Core Principles: Definitions and Terminology

According to Miller [39], interpretability in neural networks is often defined as the degree to which humans can understand or contextualize the internal reasoning of a model. Although this high-level definition captures the essence of making neural network decisions more transparent, researchers commonly break the concept down into more specific elements [16, 22]. For example, transparency focuses on inherent clarity regarding a model's structure and mechanisms, as seen in simpler algorithms like linear regression or decision trees [33]. By contrast, Doshi-Velez et al. [16] denotes that explainability involves post-hoc methods—such as feature importance analysis or local approximations—to clarify how complex "black-box" models arrive at their outputs.

According to Carvalho et al. [12], it is helpful to view interpretability as a broader umbrella that covers both transparency and explainability, while also encompassing the subjective aspect of whether end users can trust the decisions of a model. Achieving complete transparency in deep learning models is particularly challenging, given that their decision-making logic is distributed across multiple layers of nonlinear transformations [22, 33]. These architectures often achieve high accuracy but remain opaque, highlighting the tension between performance and understandability in applications where accountability is paramount (e.g., healthcare or autonomous systems). Consequently, researchers and practitioners increasingly rely on explainability techniques to offer at least partial insight into why a network makes certain predictions, even if the model itself remains complex [4]. As neural networks become more prevalent in safety-critical or ethically sensitive domains, interpretability strategies—spanning from intrinsic transparency to post hoc explanations—are critical for fostering user trust and ensuring that model outputs align with societal and regulatory expectations.

1.2 Foundational Theories

As neural networks grow increasingly complex, it becomes insufficient to rely solely on qualitative assessments, such as visualizations or case studies, to gauge interpretability [12, 56]. While these qualitative methods offer valuable insights into how a model processes information, they often lack the comprehensive, objective, and scalable nature required in high-stakes domains like healthcare and finance [40]. In these environments, continuous monitoring and comparison of interpretability are essential for ensuring trust, fairness, and accountability [3]. Quantitative metrics address this need by offering more standardized ways to evaluate interpretability, enabling consistent cross-model comparisons. They measure various facets of interpretability, including how much information is needed to explain a decision or how well a simplified model can approximate a more complex one. For instance, some metrics focus on the number of features required to justify a prediction, while others assess how closely a surrogate model replicates the original network's decision-making process [4].

The impact of these metrics is most evident in deep neural networks, which tend to function as "black-boxes" when viewed purely through introspection or basic visual analysis [56]. By quantifying interpretability, researchers can more readily identify and mitigate hidden biases or erratic behaviors—challenges especially critical in safety-sensitive applications. Additionally, such metrics facilitate the creation of benchmarks for interpretability, enabling AI systems to become more transparent and trustworthy without necessarily sacrificing performance [3]. In high-risk settings, the systematic application of quantitative interpretability measures thus becomes paramount for ensuring that neural network–driven decisions remain robust and reliable.

Alongside these quantitative considerations, an equally important dimension of interpretability is rooted in cognitive science [34]. Understanding how people process and internalize explanations helps shape interpretability metrics that are both mathematically sound and intuitively meaningful. Kulesza et al. [30] suggest that humans often prefer explanations emphasizing key causal factors over exhaustive detail, aligning with the concept of "cognitive chunks"—the manageable information units we use to form mental models. By incorporating these cognitive principles, it is possible to design metrics that more closely mirror how human users naturally seek out explanations. For example, methods assessing whether an AI-generated justification matches human-style causal reasoning—or how effectively it highlights critical comparisons (e.g., "why this outcome rather than another?")—can lead to explanations that feel more accessible and relevant.

Recent works in cognitively inspired interpretability aims to bridge these human and computational perspectives [26]. One emerging approach, known as cognitive grounding, structures AI explanations around existing human knowledge and mental schemas [9]. This means that evaluative criteria go beyond simply measuring accuracy or simplicity; they also gauge how well the explanation resonates with user expectations. Moreover, insights from visual attention research have prompted refinements to saliency map techniques, making them more reflective of how humans actually perceive and prioritize visual information [25]. By drawing on cognitive psychology and user-centric design principles, quantitative interpretability metrics can be adapted to produce explanations that are not only rigorous in their attribution of features but also well-aligned with human reasoning processes. This interdisciplinary approach promises to enhance the transparency, usability, and overall trustworthiness of neural networks—particularly in high-stakes applications where the clarity of a model's decisions can be as important as its accuracy.

1.2.1 Formal Approaches to Interpretability

As neural networks continue to be deployed in critical domains such as healthcare, finance, and autonomous systems, there is a growing demand for rigorous, mathematically grounded approaches to interpretability. Formal methods offer a structured and verifiable means of understanding how models arrive at their decisions, ensuring that explanations are not only intuitive but also provably correct [8]. Unlike heuristic-based techniques, formal approaches leverage logical reasoning, theorem proving, and mathematical abstraction to analyze and verify interpretability metrics. These methods are particularly valuable in high-stakes applications where interpretability must be trustworthy, reproducible, and resistant to adversarial manipulation.

Verification and Mathematical Rigorousness: Formal interpretability approaches often focus on model verification, ensuring that explanations hold under a broad range of conditions [17]. For instance, logic-based reasoning frameworks can be used to express and verify interpretability constraints. Consider a neural network tasked with medical diagnosis: a formal verification system can confirm that specific feature attributions (e.g., a symptom’s relevance to a disease classification) remain consistent across different patient cohorts. This ensures that the model’s decision-making process is not influenced by unintended biases or spurious correlations.

Structured Frameworks and Compositionality: One of the key challenges in interpretability is dealing with the compositional complexity of deep networks. Formal approaches aim to break down the network into interpretable submodules, where each layer or function can be independently analyzed. Automated reasoning tools, such as symbolic execution and abstract interpretation, help decompose the learned representations, making it easier to derive human-interpretable rules that govern the model’s behavior [20]. These methods facilitate a deeper understanding of layer-wise transformations and hierarchical feature extraction, which are often obscured in conventional deep learning pipelines.

Integration of Formal Methods with Existing Metrics: While many interpretability techniques focus on post-hoc analysis—such as feature attribution methods like SHAP or LIME—formal approaches embed interpretability constraints directly into the model design, ensuring that explainability is not merely an afterthought but an inherent property of the system [46]. One key direction in this space involves constraint-based interpretability, where domain-specific rules, such as monotonicity constraints in financial risk assessment models, are encoded into the network structure. This approach ensures compliance with human intuition and regulatory requirements while maintaining predictive accuracy. Another promising avenue integrates symbolic and rule-based reasoning to represent neural network decision functions as logical expressions or decision rules, making the model’s behavior more transparent and verifiable. This method aligns particularly well with explainability needs in domains where decisions must be explicitly justified, such as healthcare diagnostics or legal applications. Additionally, probabilistic formalism provides a complementary perspective by employing Bayesian inference and probabilistic graphical models to quantify uncertainty in model explanations. These approaches allow researchers to better assess the reliability of interpretability outputs, ensuring that explanations remain robust even under data shifts or adversarial perturbations.

Formalizing Interpretability through Mixed-Integer Programming: A foundational perspective on interpretability is presented by Aftabi et al. [2], where neural networks are formulated as mixed-integer programs (MIPs) to enable a structured, optimization-based approach to explainability. This formulation captures the piecewise linear characteristics of ReLU activations and other network components, allowing researchers to systematically analyze a model’s decision boundaries and feature interactions. By articulating network behavior through a set of linear constraints and binary variables, this method provides a rigorous mathematical framework for interpretability, complementing traditional metrics such as fidelity and complexity while also facilitating verifiable robustness and fairness assessments.

Unlike empirical interpretability methods that rely on post-hoc approximations, MIP-based approaches provide exact insights into network decision-making. By precisely characterizing how neural networks partition input space, this methodology offers a deeper understanding of model predictions and the factors influencing them. This capability is particularly relevant in regulated or high-stakes applications where explainability is crucial for ensuring transparency, accountability, and compliance with ethical standards. Furthermore, by integrating formal verification techniques with cognitively informed interpretability metrics, MIP-based methods extend the scope of what can be reliably understood about neural network models. In doing so, they reinforce the role of optimization techniques not just as tools for improving performance but also as essential mechanisms for achieving greater interpretability in complex AI systems.

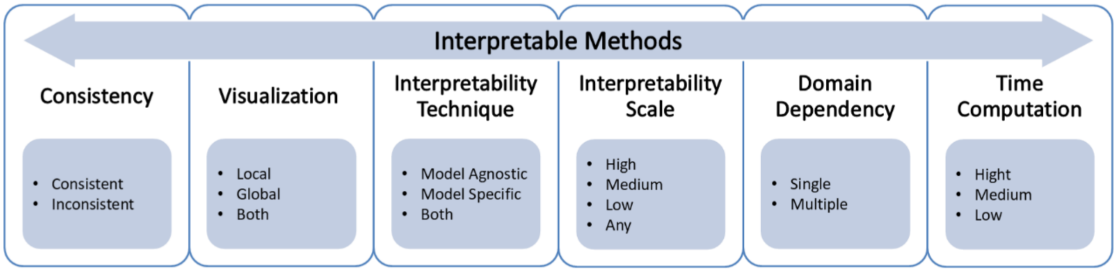

Figure 1. Classification of Interpretable Methods Across Key Dimensions.

2. Measuring Interpretability: Framework and Metrics

2.1 Types of Quantitative Interpretability Metrics

Quantitative interpretability metrics are essential for evaluating the comprehensibility and effectiveness of machine learning models. These metrics can be categorized into several types, each serving a specific purpose in assessing how interpretable a model is. Interpretability techniques can be broadly classified into two categories: intrinsic interpretability and post-hoc interpretability. Intrinsic interpretability involves using inherently understandable models, such as linear models or decision trees, which provide clear insights into the relationships between inputs and outputs [12]. Post-hoc interpretability, on the other hand, applies to complex models where interpretability methods are used after the model has been developed to explain its behavior and predictions [42]. Figure 1 summarizes key characteristics of different interpretability methods, emphasizing their complexity, domain dependency, and computational requirements.

Feature Importance: Feature importance metrics evaluate the relative impact of different features on the predictions made by a model. They help identify which features significantly influence outcomes and can guide decisions on data preprocessing or feature selection. Common methods for calculating feature importance include Gini importance and permutation importance, which quantify each feature’s contribution to the model’s predictive accuracy [12]. These metrics are crucial for understanding model behavior and enhancing interpretability.

Visualization Techniques: According to [56], visualization techniques are employed to present the relationships between features and model predictions clearly. Tools like decision trees, heatmaps, and partial dependence plots provide intuitive insights into how features interact and contribute to outcomes [29, 42]. For instance, decision trees offer a straightforward representation of decision paths based on feature values, while heatmaps can illustrate feature activation and attention in deep learning models. Such visual aids make it easier for users to grasp complex relationships within the data.

Explanation Types: Explanations can be classified into local and global types, which cater to different interpretability needs. Local explanations pertain to individual predictions, while global explanations provide insights into the model as a whole [53]. Different explanation methods yield various types of outputs, such as rule lists or graphical representations, allowing for diverse approaches to understanding model decisions.

Evaluation Metrics for Interpretability: When assessing the interpretability of a model or its explanations, specific evaluation metrics can be employed. These may include intuitiveness, determinism, generalizability, and faithfulness [36]. Intuitiveness ensures that the metric is easily understandable, while determinism guarantees consistent outputs given the same inputs. Generalizability indicates the metric’s applicability across different models, and faithfulness reflects the alignment of the metric with the model’s decision-making process.

Consistency: The consistency of interpretations across various attribution methods is also an important consideration. Metrics that evaluate the goodness of attribution maps—both at the instance level and globally—can be employed to judge whether different explanation methods yield coherent results. This approach enhances trust in the interpretability of machine learning models by ensuring that explanations are consistent and reliable [43]. By utilizing these diverse types of quantitative interpretability metrics, researchers and practitioners can effectively assess and enhance the interpretability of machine learning models, fostering trust and understanding among stakeholders.

2.2 Model-Faithful Explanations (Fidelity)

Fidelity metrics are essential for evaluating interpretability in neural networks, as they measure how accurately an explanation reflects the model's true decision-making process [28]. Often referred to as faithfulness, fidelity indicates whether an interpretability method faithfully captures the logic underlying the model's predictions rather than presenting a distorted or oversimplified view. High-fidelity explanations are especially crucial in high-stakes settings, such as healthcare and finance, where user trust in the model's reasoning can significantly impact outcomes.

One prominent example illustrating the importance of fidelity is LIME (Locally Interpretable Model-Agnostic Explanations) [44]. LIME uses a local linear surrogate model to approximate the predictions of a more complex "black-box" model for specific data points [44]. The degree to which this surrogate's predictions align with those of the original model indicates how faithful its explanations are. When alignment is high, LIME accurately mirrors the original model's decision-making logic; if alignment is low, the resulting explanations may mislead users into believing the model relies on factors that do not actually drive its predictions. Consequently, fidelity metrics are indispensable for determining whether an explanation genuinely reflects the inner workings of a neural network. Without them, explanations can easily become misleading, eroding user confidence and potentially causing harmful misinterpretations in mission-critical applications.

2.3 Complexity and Comprehensibility

Complexity metrics gauge how easily humans can understand a model's explanations by quantifying factors such as the number of features involved, the depth of decision paths, and the overall cognitive load on the user [18]. Generally, simpler explanations are more desirable, as they reduce the mental effort required to grasp how the model generates its outputs. For example, a decision tree's interpretability can be assessed by examining its maximum depth—shorter trees with fewer branching features tend to be more transparent. Similarly, rule-based models can be evaluated according to the number and length of their rules; larger or more intricate rule sets typically impose greater cognitive demands.

In neural networks, complexity measures often center on identifying the key input features that significantly influence a model's predictions [13]. Feature attribution techniques, such as Shapley values or Integrated Gradients [50], help determine whether the network relies on a small, interpretable subset of inputs or a broader, more opaque interaction of features. Although highly complex models may offer superior accuracy, their opacity can undermine trust—particularly in high-stakes domains like healthcare or finance, where stakeholders require a clear understanding of automated decisions. By employing complexity metrics, practitioners can balance model performance with interpretability, ensuring that AI-driven outcomes remain both efficient and comprehensible for end users.

2.4 Reliability and Robustness

Reliability and robustness are critical aspects of interpretability, particularly in high-stakes domains where fluctuating or inconsistent explanations can quickly erode confidence in AI-driven systems [5]. Two complementary dimensions of robustness include how explanations respond to small input changes and how they adapt over time as models learn continuously.

Stability and Consistency: Stability metrics evaluate whether minor modifications in input data lead to dramatically different explanations [12]. In high-stakes scenarios, such as healthcare diagnoses or financial assessments, an interpretability method that yields vastly different explanations for nearly identical inputs can undermine trust and suggest that the model is overly sensitive to small perturbations. To measure stability, researchers often examine the variance in explanations when input features are slightly altered, using techniques such as adversarial tests or sensitivity analysis [12]. Consistently high variance can indicate unreliable feature attribution or latent vulnerabilities in the model. By embedding stability and consistency checks into an interpretability framework, practitioners can ensure explanations remain coherent and dependable, even as underlying data shifts incrementally.

Temporal Consistency: As neural networks are increasingly deployed in dynamic environments, where models continuously learn and adapt, a second layer of robustness revolves around the stability of explanations over time. Temporal consistency metrics track whether a model’s explanations remain coherent across different iterations or updates—an especially important concern in online learning or frequently retrained systems [27]. One approach, the Temporal Stability Index (TSI), measures how much feature importance fluctuates across successive model versions [22]. A low TSI typically indicates stable decision-making criteria, whereas a high TSI may signal significant shifts in feature importance over time. Similarly, the Explanation Drift Rate (EDR) quantifies how quickly explanations for similar inputs diverge across model updates [22]. Sudden spikes in EDR can reveal concept drift or structural changes in how the network processes information.

By incorporating both stability and temporal consistency assessments, researchers and practitioners can better detect when a model’s interpretation pipeline becomes susceptible to noise or begins to rely on newly introduced but potentially spurious patterns. This dual focus on small-scale perturbations and longitudinal changes is vital for maintaining trustworthy, real-world AI systems. In contexts ranging from personalized healthcare to automated trading, robust interpretability measures not only clarify current model behavior but also help anticipate future shifts—ensuring that evolving neural networks remain transparent, reliable, and aligned with user expectations.

2.5 Feature Attribution and Sensitivity

Feature attribution and sensitivity metrics focus on how variations in input features influence a model's predictions, as well as how these effects are reflected in explanatory outputs [36]. By quantifying each feature's contribution, these metrics not only highlight the most significant drivers of a model's decisions but also reveal how sensitive the model is to changes in individual inputs. In doing so, they help ensure the model behaves consistently and aligns with domain-specific knowledge and user expectations.

One well-known approach to feature attribution is based on Shapley values, which originate from cooperative game theory and provide a theoretically grounded method for distributing the contribution of each feature to the final prediction [46, 50]. A feature's sensitivity can be assessed by examining the fluctuation in its Shapley value when the underlying input changes. Beyond Shapley values, methods such as Integrated Gradients and LIME offer additional frameworks for assigning feature importance, each with distinct strengths and limitations [44]. These attribution and sensitivity analyses are particularly critical in high-dimensional or complex models, where the roles of individual features are not immediately evident. By systematically applying these metrics, researchers and practitioners can uncover potential biases, detect inconsistencies in a model's decision-making process, and ultimately enhance the transparency and trustworthiness of AI-driven applications.

3. Evaluating Interpretability in Practice

3.1 Empirical Validation and Case Examples

Empirical validation is crucial for determining how effectively quantitative interpretability metrics capture the decision-making processes of neural networks. In a typical experimental setup, researchers train a network on a benchmark dataset—such as MNIST, CIFAR-10, or one from the UCI Machine Learning Repository—using architectures like convolutional or recurrent neural networks. They then apply interpretability methods, for example LIME or SHAP, to generate explanations. These explanations are subsequently evaluated using a combination of metrics: fidelity measures how closely the surrogate explanations align with the model's true decision-making process, complexity gauges the simplicity and comprehensibility of the generated explanations, and robustness and stability tests whether minor changes in the input data lead to drastically different explanations. By employing such a framework, researchers can systematically compare how various metrics perform under controlled conditions.

In addition to these standardized evaluations, case studies in real-world applications provide critical qualitative insights into how well interpretability metrics translate to practical settings. In healthcare, for instance, neural networks trained on electronic health records (EHRs) to predict patient outcomes have been examined through metrics like fidelity, complexity, and robustness [11]. Such metrics help verify whether explanations accurately reflect the model's reasoning, determine whether those explanations are clear enough for clinicians to interpret and act upon, and ensure that minor variations in patient data do not significantly alter the model's justifications. This approach reinforces reliability in clinical decision-making and bolsters trust in AI-driven diagnoses.

Similarly, in finance, interpretability metrics are applied to credit scoring models to validate transparency and fairness in automated lending decisions [36]. Fidelity analyses check if predictions are grounded in meaningful features such as income or credit history, while attribution and sensitivity methods reveal whether irrelevant or biased inputs play a role in the final outcome. Stability measures, in turn, confirm that minor changes in an applicant's data do not unfairly alter creditworthiness assessments. By demonstrating accountability and ensuring consistent treatment across similar cases, these techniques promote fairer financial services. In general, empirical testing, coupled with domain-specific case studies, underscores the importance of quantitative interpretability metrics for validating and refining AI models across diverse contexts. By integrating both quantitative benchmarks and practical insights, researchers and practitioners can better guarantee that neural networks remain accurate, transparent, and reliable in real-world applications.

3.2 Comparative Analysis

Each interpretability metric offers distinct advantages and limitations, making certain methods more suitable than others depending on the application context. Fidelity metrics, for instance, excel in verifying that explanations truly reflect a model's decision-making process—an aspect critical in high-stakes scenarios like healthcare diagnostics or autonomous systems, where precision and trust are paramount. However, focusing on fidelity alone may not capture how easily humans can understand an explanation. In situations where usability and intuitive clarity are key—such as consumer-facing AI systems or clinical decision-support tools—complexity metrics become essential for assessing whether an explanation is sufficiently simple to be grasped by non-experts [12].

Robustness and stability metrics, meanwhile, address the need for consistent, repeatable outputs. These measures are particularly relevant in regulated or legal environments, where decisions must be justifiable across similar cases. Nevertheless, robustness alone may not clarify the underlying reasons for a given output. In domains such as personalized medicine, credit scoring, or targeted marketing, sensitivity and attribution methods, including Shapley values, play a pivotal role. They determine which features most significantly influence predictions, ensuring that decisions align with domain-specific knowledge and fairness requirements [5].

Choosing the right interpretability metric or combination thereof ultimately depends on each task's specific requirements. Sensitive healthcare applications, for example, may benefit from pairing fidelity metrics with complexity assessments to ensure both accuracy and comprehensibility in clinical decision-making. On the other hand, financial contexts—like credit scoring or fraud detection—may demand robust explanations that withstand regulatory scrutiny, making stability and attribution measures more critical. By aligning interpretability strategies with the nuances of each domain, practitioners can maintain transparent, trustworthy, and domain-tailored AI solutions across diverse real-world scenarios.

3.3 Tools and Software Ecosystem

A variety of tools and frameworks facilitate the assessment of interpretability using quantitative measures, providing practical methods for generating and evaluating explanations. One of the most widely used options is LIME, which explains individual predictions by fitting a simple, interpretable surrogate model around the local neighborhood of the original complex model. LIME is particularly valuable for assessing fidelity and complexity, as it allows researchers to determine how closely the surrogate's behavior matches that of the underlying model and how easily its explanations can be understood [44].

Another popular framework is SHAP, which leverages Shapley values to provide accurate, consistent feature attributions [36]. SHAP proves especially effective in evaluating sensitivity and attribution metrics, thanks to its unified approach for quantifying the importance of each input feature across different types of models. Beyond LIME and SHAP, libraries such as Captum and Alibi also offer extensive toolsets for various interpretability techniques. Captum is designed for PyTorch and includes methods like Integrated Gradients and Feature Ablation, both of which are useful for assessing robustness and stability. Alibi, on the other hand, is an open-source Python library that supports counterfactual explanations and adversarial detection, contributing additional insights into model stability and reliability. By integrating these tools into their workflows, researchers and practitioners can establish consistent benchmarks for interpretability across diverse neural network architectures and application domains. This shared ecosystem ensures that AI models remain transparent and interpretable while also aligning with the operational needs and constraints of real-world use cases.

4. Domain-Specific Perspectives

Neural networks are increasingly adopted in high-stakes fields, making interpretability a vital concern for ensuring trust, compliance, and safety. In medical diagnostics, for example, clinicians must validate AI-driven predictions before making decisions that can significantly impact patient outcomes [19, 49]. Fidelity metrics help confirm that the factors influencing diagnostic predictions—such as specific symptoms or test results—are indeed the ones driving the model's outputs, thereby preventing reliance on spurious correlations. Complexity metrics also play a crucial role, since overly detailed explanations can slow down urgent clinical decisions. By contrast, simpler and more transparent explanations support swift, accurate judgment. At the same time, robustness and stability measures guard against erratic responses to minor perturbations in patient data, thus promoting consistent, trustworthy performance in healthcare settings.

In the financial sector, interpretability is central to risk assessment, fraud detection, and automated decision-making, especially given strict regulatory requirements [6, 14]. Fidelity metrics verify that credit scoring models rely on legitimate factors such as credit history or income rather than irrelevant or unethical criteria. Complexity measures ensure that stakeholders—like loan officers and auditors—can readily understand and scrutinize model explanations. Sensitivity and attribution metrics, including Shapley values, help identify key features driving fraud detection algorithms, enabling financial institutions to pinpoint suspicious transactions and justify their decisions [32]. Robustness assessments further maintain consistent treatment of applicants with similar financial profiles, thereby reducing bias and enhancing fairness in lending and other financial service [16].

Beyond healthcare and finance, interpretability metrics play a key role in cybersecurity, where explainability is essential for identifying potential adversarial threats and improving model robustness against attacks [41]. Understanding the reasoning behind an anomaly detection system's alert, for instance, allows security teams to distinguish between genuine threats and false positives more effectively [1]. Fidelity metrics ensure that the model's explanations align with the underlying security threats, while stability assessments confirm that the detection logic remains consistent under minor variations in attack patterns.

In autonomous systems, trust in decision-making is paramount, particularly in self-driving cars, drones, and robotic automation [47, 55]. Interpretability techniques help ensure that AI-driven navigation systems make decisions based on legitimate environmental cues rather than spurious correlations. Complexity reduction methods allow engineers to analyze decision pathways, ensuring that models remain comprehensible without sacrificing critical functionality. Robustness evaluations further contribute by detecting whether slight changes in inputs—such as lighting variations or sensor noise—cause disproportionate shifts in model behavior. By maintaining transparency in decision-making, interpretability metrics improve both safety and public acceptance of autonomous systems.

Energy systems also benefit from explainability, particularly in optimizing power consumption, forecasting demand, and integrating renewable energy sources [10, 15, 37]. Neural networks used in energy management must provide clear justifications for their predictions and recommendations, ensuring that power distribution models remain both efficient and interpretable. Sensitivity analysis helps identify the most influential variables driving energy predictions, ensuring that optimization strategies align with real-world constraints. By improving interpretability, AI models in the energy sector can provide more reliable guidance for balancing grid demand and supply, ultimately supporting the transition to sustainable energy systems.

Industrial AI applications, including predictive maintenance and manufacturing process optimization, similarly rely on interpretability metrics to ensure transparent decision-making [2, 24]. When AI models predict equipment failures or recommend process adjustments, clear explanations are necessary for engineers and operators to trust and act on these insights. Feature attribution techniques allow practitioners to pinpoint the critical variables influencing a failure prediction, ensuring that maintenance strategies are both data-driven and actionable. Complexity metrics further ensure that the generated explanations remain practical and accessible to non-specialists, facilitating effective decision-making on the factory floor.

Table 1 provides a comparative framework summarizing the strengths, limitations, and applications of key interpretability metrics across various domains. This overview consolidates the different dimensions of interpretability, offering a structured reference for researchers and practitioners seeking to evaluate or implement these metrics in real-world AI applications. Across these diverse sectors, interpretability metrics form a crucial foundation for building AI models that meet ethical, legal, and social expectations. Ensuring that explanations are accurate, transparent, and consistent enhances user trust, simplifies compliance, and contributes to safer, more accountable decision-making. As AI technologies evolve and become further integrated into complex environments, more advanced interpretability metrics will be needed to address emerging challenges. A balanced approach—one that does not sacrifice performance for understandability—remains essential, and standardizing interpretability assessment across industries will be key to fostering confidence and facilitating regulatory alignment. By prioritizing interpretability in AI development, researchers and practitioners can achieve solutions that maintain high accuracy while also supporting transparency, fairness, and user trust in real-world applications.

Table 1. Comparative Framework for Interpretability Metrics.

Metric Name	**Strengths**	**Weaknesses**	**High-Level Application**	**General Advantage /Limitation**
Fidelity	Accurately reflects model behavior; ensures explanations align with true decision-making logic.	May not capture user-comprehensible insights; requires validation with model outputs.	Healthcare, finance, autonomous systems.	High fidelity ensures trustworthiness, but complex models may still be difficult to interpret.
Complexity	Enhances human understanding;	Can oversimplify model behavior,	Decision-support	Strikes a balance between

Metric Name	Strengths	Weaknesses	High-Level Application	General Advantage /Limitation
	evaluates the cognitive load required to comprehend explanations.	potentially leading to loss of important details.	systems, legal AI, consumer-facing applications.	transparency and usability; overly simplistic models may lack expressiveness.
Robustness	Measures stability of explanations across small perturbations in input data.	High robustness may conflict with model adaptability; does not directly assess accuracy.	Adversarial defenses, safety-critical applications.	Ensures consistency in explanations, but may limit responsiveness to meaningful data changes.
Sensitivity	Evaluates how variations in inputs affect explanations and feature importance.	Highly sensitive explanations may lack reliability; unstable models can degrade interpretability.	Fraud detection, risk assessment, scientific modeling.	Provides insight into decision boundaries, but excessive sensitivity can introduce instability.
Temporal Consistency	Tracks whether explanations remain stable over time and across model updates.	May not detect slow concept drift; requires careful selection of evaluation timescales.	Online learning systems, adaptive AI, time-sensitive decisions.	Helps ensure long-term trust in AI predictions, but may hinder model flexibility.

5. Cross-Cutting Challenges and Future Directions

5.1 Data Quality, Bias, and Ethical Concerns

Data quality is fundamental to both the development and accurate assessment of interpretability metrics in neural networks. When training data are skewed or incomplete, resulting models often produce misleading explanations that fail to reflect their true decision-making process. A particularly pressing issue arises when the datasets used for interpretability evaluation carry inherent biases traced to historical inequalities, sampling errors, or flawed data collection methodologies [38]. These biases can become embedded in a model's logic, causing it to systematically favor or disadvantage certain groups. When interpretability methods are subsequently applied, they may unintentionally legitimize these imbalances by highlighting features that stem from the biased data, thereby reinforcing existing inequities. In high-stakes contexts such as healthcare or finance, this dynamic can lead to ethically questionable or even discriminatory outcomes.

Addressing these concerns requires a holistic approach to data collection, preprocessing, and validation, ensuring that the final dataset is both representative and free from systematic bias. Fairness-aware interpretability measures also play an essential role by detecting and mitigating bias at multiple levels, from training data to model outputs. Techniques such as counterfactual explanations, adversarial testing, and fairness constraints can pinpoint where biased features disproportionately shape decisions, making it easier to implement corrective measures that enhance both interpretability and fairness [21]. As neural networks continue to expand their reach, developing bias-aware interpretability metrics will be crucial for maintaining transparency, ethical accountability, and alignment with societal values.

5.2 Scalability and Complex Architectures

As neural networks become increasingly intricate, it becomes far more challenging to apply interpretability metrics that can reliably capture their decision-making processes. Convolutional neural networks (CNNs), designed to extract spatial hierarchies from large numbers of parameters, introduce complexities that do not necessarily arise in simpler models. Likewise, recurrent neural networks (RNNs) and advanced variants such as long short-term memory networks (LSTMs) and gated recurrent units (GRUs) add another layer of difficulty by incorporating temporal dependencies. Methods that work well for basic feed-forward architectures may struggle when tasked with explaining deeper or more specialized structures [16].

Ensuring that interpretability metrics remain scalable thus requires techniques capable of handling high-dimensional feature spaces and the unique dependencies characterizing different network types. A metric tailored for CNNs, for instance, often focuses on spatial feature importance and may not effectively reveal critical temporal patterns in RNN-based models. Conversely, methods originally developed for fully connected networks might fail to account for the sequential nature of time-series data. Attention-based explanations, layer-wise relevance propagation, and gradient-based attribution methods represent some of the current strategies for extending interpretability to deeper and more diverse architectures. However, generalizability remains a major concern, since a metric validated on one dataset may yield inconsistent results when applied to new domains or data distributions [5].

Future advances in interpretability will likely hinge on creating standardized frameworks that adapt across various neural architectures while retaining consistency, reliability, and real-world applicability. As these networks continue to expand in both scope and depth, robustly scalable metrics will be indispensable for ensuring that high-performing models also maintain transparent, interpretable decision-making processes.

5.3 Temporal Consistency: Challenges and Applications

A critical challenge in ensuring robust interpretability is that model explanations must remain coherent and trustworthy even as the underlying models adapt to new data or changing conditions. This notion of temporal consistency focuses on how explanations evolve over time and how the logic behind predictions may shift due to retraining, online learning, or external factors [5]. Determining the appropriate time scale for evaluating these changes poses a fundamental dilemma: models must be allowed to adapt naturally while still preserving stable, meaningful explanations that stakeholders can trust. Temporal consistency metrics become especially important in dynamic environments such as fraud detection or recommender systems, where user behavior and data distributions can shift rapidly [22, 27]. In these contexts, models must be recalibrated to incorporate fresh data without compromising their interpretability. As a result, temporal consistency metrics must be calibrated to strike a balance between enabling essential model updates and identifying any unintended or unstable changes that could undermine user confidence. High-stakes applications in healthcare and finance further underscore the value of temporal consistency. For instance, in clinical diagnostics, maintaining stable interpretations over a patient’s care pathway is critical for clinician acceptance and safe patient outcomes [19, 49]. Likewise, in financial risk assessment—where macroeconomic trends and market volatility can shift rapidly—temporal consistency metrics help ensure that model outputs remain both transparent and reliable over extended periods [6, 14]. By integrating temporal consistency assessments into broader interpretability frameworks, practitioners can track how a model's decision rationale progresses as conditions evolve. This longitudinal perspective not only detects when a model might begin relying on spurious correlations or experiencing concept drift but also ensures that explanations continue to be trustworthy and actionable, even in the face of ongoing changes in data and domain requirements.

5.4 Towards Inherently Interpretable Networks

Embedding interpretability metrics into the model development lifecycle is increasingly viewed as essential for ensuring that neural networks are transparent from the outset, rather than treating interpretability as a post hoc concern. Traditional workflows often emphasize accuracy at the expense of transparency, leading to highly complex models that excel at predictive performance yet are difficult to understand [45]. In response, recent research has shifted toward designing inherently interpretable models, where explainability is treated as a core design criterion. This approach includes integrating interpretability metrics directly into the training process so that networks can optimize both predictive accuracy and transparency simultaneously [35, 51, 52].

A variety of strategies facilitate this goal. Limiting model complexity or imposing constraints on feature usage ensures that models do not grow unnecessarily large or obscure. Intrinsically interpretable architectures such as decision trees, rule-based models, and linear models provide built-in transparency, while still achieving strong performance in many domains. Additionally, regularization methods—ranging from sparsity-inducing penalties to attention mechanisms and disentangled representations—can further enhance interpretability by constraining parameter space. Models that generate human-readable explanations alongside predictions, such as self-explaining neural networks (SENN) and certain attention-based designs, represent another promising avenue for striking a balance between accuracy and understandability [4].

By weaving interpretability into every stage of model design and development, researchers and practitioners can create AI systems that are not only robust and accurate but also inherently trustworthy. This integrated perspective is especially critical in domains like healthcare, finance, and legal decision-making, where interpretability is fundamental to regulatory compliance, fairness, and user acceptance. Moving forward, adopting inherently interpretable architectures and training processes will be a key step toward building AI systems that align more seamlessly with ethical standards and human decision-making needs.

5.5 Emerging Paradigms

The field of neural network interpretability is expanding to address the limitations of existing approaches, prompting the development of new and innovative metrics. One notable advancement involves causal interpretability measures, which move beyond simple correlations to uncover the direct causal effects of individual features on model outputs [7]. By evaluating whether altering a specific feature truly changes a model's prediction, these metrics offer deeper insights into the underlying decision-making process. They are especially critical in applications such as healthcare or finance, where understanding cause-and-effect relationships helps ensure fairness, accountability, and compliance with ethical or legal standards.

Another promising direction focuses on dynamic interpretability metrics that adapt to various stages of model training and deployment [23]. Whereas many current methods provide static explanations once a model is fully trained, these newer approaches integrate interpretability assessment throughout the entire lifecycle. Developers can then receive real-time feedback on shifts in feature importance, architecture changes, and emerging patterns in the decision process. Such continuous monitoring proves particularly valuable in iterative or online learning scenarios, where models are frequently retrained or refined. By keeping interpretability in sync with evolving model parameters, stakeholders can maintain consistent levels of transparency and ensure that the explanations remain relevant over time.

Additionally, interactive interpretability metrics have begun to emerge, offering users the ability to engage with a model's explanations and provide feedback that adjusts them in real time [48]. This approach accommodates the fact that different user groups—ranging from regulators and domain experts to everyday end users—may require different levels of detail or specific perspectives on model behavior. By incorporating user feedback into interpretability, AI systems can generate context-sensitive explanations that better align with individual decision needs and expertise levels. These developments collectively signal a future where interpretability is not only more granular and accurate, but also more responsive and human-centric, opening the door to AI systems that are transparent, adaptable, and ethically responsible in their decision-making processes.

6. Conclusion

Quantitative metrics play a pivotal role in evaluating the interpretability of neural networks, which have become integral to high-stakes domains such as finance, healthcare, and autonomous systems.

This paper has reviewed a range of these metrics—fidelity, complexity, stability, and sensitivity—each providing distinct insights into how well explanations align with a model's underlying logic, how easily they can be understood, and how resilient they remain under different conditions. Despite substantial progress in developing such metrics, several critical challenges persist. Data quality, bias, and scalability remain significant obstacles, as interpretability approaches validated on simpler or cleaner datasets may not generalize effectively to more complex real-world settings. Moreover, a common practice of treating interpretability as an afterthought rather than integrating it into the model development process has limited the impact of current methods. Emerging directions—including causal interpretability and interactive metrics—show promise for bridging the gap between technical explanations and genuine human understanding. Yet fully transparent and trustworthy AI systems remain an aspirational goal. The tension between maximizing model performance and ensuring clarity, along with the need for metrics that can adapt to evolving architectures, underscores the importance of ongoing research. Overcoming these challenges is crucial for building AI models that are not only accurate but also explainable, ethical, and compliant with regulatory standards. By continuing to refine interpretability metrics and embedding them throughout the AI lifecycle, researchers and practitioners can bolster trust in AI-driven systems and ensure that neural networks operate in ways that are transparent, accountable, and aligned with societal values.

References

[1] Aftabi, N., Li, D., & Ramanan, P. (2023). A variational autoencoder framework for robust, physics-informed cyberattack recognition in industrial cyber-physical systems. Retrieved from https://arxiv.org/abs/2310.06948

[2] Aftabi, N., Moradi, N., & Mahroo, F. (2025). Feed-forward neural networks as a mixed-integer program. *Engineering with Computers*. https://doi.org/10.1007/s00366-025-02114-2

[3] Ahmed, I., Jeon, G., & Piccialli, F. (2022). From artificial intelligence to explainable artificial intelligence in industry 4.0: A survey on what, how, and where. *IEEE Transactions on Industrial Informatics*, *18*(8), 5031–5042. https://doi.org/10.1109/TII.2022.3146552

[4] Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, *99*, 101805. https://doi.org/https://doi.org/10.1016/j.inffus.2023.101805

[5] Alvarez Melis, D., & Jaakkola, T. (2018). Towards robust interpretability with self-explaining neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 31). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2018/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf

[6] Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. Retrieved from https://arxiv.org/abs/1806.08049

[7] Andreas G. F. Hoepner, A. V., David McMillan, & Simen, C. W. (2021). Significance, relevance and explainability in the machine learning age: An econometrics and financial data science perspective. *The European Journal of Finance*, *27*(1-2), 1–7. https://doi.org/10.1080/1351847X.2020.1847725

[8] Baiardi, A., & Naghi, A. A. (2024). The value added of machine learning to causal inference: Evidence from revisited studies. *The Econometrics Journal*, *27*(2), 213–234. https://doi.org/10.1093/ectj/utae004

[9] Barceló, P., Monet, M., Pérez, J., & Subercaseaux, B. (2020). Model interpretability through the lens of computational complexity. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 15487–15498). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2020/file/b1adda14824f50ef24ff1c05bb66faf3-Paper.pdf

[10] Barsalou, L. W. (2010). Grounded cognition: Past, present, and future. *Topics in Cognitive Science*, *2*(4), 716–724. https://doi.org/https://doi.org/10.1111/j.1756-8765.2010.01115.x

[11] Baur, L., Ditschuneit, K., Schambach, M., Kaymakci, C., Wollmann, T., & Sauer, A. (2024). Explainability and interpretability in electric load forecasting using machine learning techniques – a review. *Energy and AI*, *16*, 100358. https://doi.org/https://doi.org/10.1016/j.egyai.2024.100358

[12] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for HealthCare: Predicting pneumonia risk and hospital 30-day readmission. In (pp. 1721–1730). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2783258.2788613

[13] Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8). https://doi.org/10.3390/electronics8080832

[14] Chakraborty, S., Tomsett, R., Raghavendra, R., Harborne, D., Alzantot, M., Cerutti, F., Srivastava, M., Preece, A., Julier, S., Rao, R. M., Kelley, T. D., Braines, D., Sensoy, M., Willis, C. J., & Gurram, P. (2017). Interpretability of deep learning models: A survey of results. In *2017 IEEE SmartWorld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, internet of people and smart city innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)* (pp. 1–6). https://doi.org/10.1109/UIC-ATC.2017.8397411

[15] Chen, C., Lin, K., Rudin, C., Shaposhnik, Y., Wang, S., & Wang, T. (2022). A holistic approach to interpretability in financial lending: Models, visualizations, and summary-explanations. *Decision Support Systems*, *152*, 113647. https://doi.org/https://doi.org/10.1016/j.dss.2021.113647

[16] Dong, W., Chen, X., & Yang, Q. (2022). Data-driven scenario generation of renewable energy production based on controllable generative adversarial networks with interpretability. *Applied Energy*, *308*, 118387. https://doi.org/https://doi.org/10.1016/j.apenergy.2021.118387

[17] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. Retrieved from https://arxiv.org/abs/1702.08608

[18] Dvijotham, K., Garnelo, M., Fawzi, A., & Kohli, P. (2018). Verification of deep probabilistic models. Retrieved from https://arxiv.org/abs/1812.02795

[19] Ehret, K., Blumenthal-Dramé, A., Bentz, C., & Berdicevskis, A. (2021). Meaning and measures: Interpreting and evaluating complexity metrics. *Frontiers in Communication*, *6*. https://doi.org/10.3389/fcomm.2021.640510

[20] ElShawi, R., Sherif, Y., Al-Mallah, M., & Sakr, S. (2021). Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Computational Intelligence*, *37*(4), 1633–1650. https://doi.org/https://doi.org/10.1111/coin.12410

[21] Fang, B., Yang, E., & Xie, F. (2020). Symbolic techniques for deep learning: Challenges and opportunities. Retrieved from https://arxiv.org/abs/2010.02727

[22] Ferrara, E. (2024). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci*, *6*(1). https://doi.org/10.3390/sci6010003

[23] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th international conference on data science and advanced analytics (DSAA)* (pp. 80–89). https://doi.org/10.1109/DSAA.2018.00018

[24] Glanois, C., Weng, P., Zimmer, M., Li, D., Yang, T., Hao, J., & Liu, W. (2024). A survey on interpretable reinforcement learning. *Machine Learning*, *113*(8), 5847–5890. https://doi.org/10.1007/s10994-024-06543-w

[25] Goldman, C. V., Baltaxe, M., Chakraborty, D., Arinez, J., & Diaz, C. E. (2023). Interpreting learning models in manufacturing processes: Towards explainable AI methods to improve trust in classifier predictions. *Journal of Industrial Information Integration*, *33*, 100439. https://doi.org/https://doi.org/10.1016/j.jii.2023.100439

[26] Hassanin, M., Anwar, S., Radwan, I., Khan, F. S., & Mian, A. (2024). Visual attention methods in deep learning: An in-depth survey. *Information Fusion*, *108*, 102417. https://doi.org/https://doi.org/10.1016/j.inffus.2024.102417

[27] He, Z., Achterberg, J., Collins, K., Nejad, K., Akarca, D., Yang, Y., Gurnee, W., Sucholutsky, I., Tang, Y., Ianov, R., Ogden, G., Li, C., Sandbrink, K., Casper, S., Ivanova, A., & Lindsay, G. W. (2024). Multilevel interpretability of artificial neural networks: Leveraging framework and methods from neuroscience. Retrieved from https://arxiv.org/abs/2408.12664

[28] Ismail, A. A., Gunady, M., Corrada Bravo, H., & Feizi, S. (2020). Benchmarking deep learning interpretability in time series predictions. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural*

information processing systems (Vol. 33, pp. 6441–6452). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2020/file/47a3893cc405396a5c30d91320572d6d-Paper.pdf

[29] Jacovi, A., & Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? Retrieved from https://arxiv.org/abs/2004.03685

[30] Kolyshkina, I., & Simoff, S. (2021). Interpretability of machine learning solutions in public healthcare: The CRISP-ML approach. *Frontiers in Big Data*, *4*. https://doi.org/10.3389/fdata.2021.660206

[31] Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., & Wong, W.-K. (2013). Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE symposium on visual languages and human centric computing* (pp. 3–10). https://doi.org/10.1109/VLHCC.2013.6645235

[32] Liang, Y., Machado, M. C., Talvitie, E., & Bowling, M. (2016). State of the art control of atari games using shallow reinforcement learning. Retrieved from https://arxiv.org/abs/1512.01563

[33] Lin, K., & Gao, Y. (2022). Model interpretability of financial fraud detection by group SHAP. *Expert Systems with Applications*, *210*, 118354. https://doi.org/https://doi.org/10.1016/j.eswa.2022.118354

[34] Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31–57.

[35] Lombrozo, T. (2009). Explanation and categorization: How "why?" Informs "what?" *Cognition*, *110*(2), 248–253. https://doi.org/https://doi.org/10.1016/j.cognition.2008.10.007

[36] Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Ser, J. D., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Páez, A., Samek, W., Schneider, J., Speith, T., & Stumpf, S. (2024). Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, *106*, 102301. https://doi.org/https://doi.org/10.1016/j.inffus.2024.102301

[37] Lundberg, S. (2017). A unified approach to interpreting model predictions. *arXiv Preprint arXiv:1705.07874*.

[38] Manfren, M., Gonzalez-Carreon, K. M., & James, P. A. B. (2024). Interpretable data-driven methods for building energy modelling—a review of critical connections and gaps. *Energies*, *17*(4). https://doi.org/10.3390/en17040881

[39] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, *54*(6). https://doi.org/10.1145/3457607

[40] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/https://doi.org/10.1016/j.artint.2018.07.007

[41] Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: A comprehensive review. *Artificial Intelligence Review*, 1–66. https://doi.org/https://doi.org/10.1007/s10462-021-10088-y

[42] Mohale, V. Z., & Obagbuwa, I. C. (2025). A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity. *Frontiers in Artificial Intelligence*, *8*. https://doi.org/10.3389/frai.2025.1526221

[43] Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., Keulen, M. van, & Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Comput. Surv.*, *55*(13s). https://doi.org/10.1145/3583558

[44] Pillai, V., & Pirsiavash, H. (2021). Explainable models with consistent interpretations. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(3), 2431–2439. https://doi.org/10.1609/aaai.v35i3.16344

[45] 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* (pp. 1135–1144). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

[46] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[47] Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2025). A perspective on explainable artificial intelligence methods: SHAP and LIME. *Advanced Intelligent Systems*, *7*(1), 2400304. https://doi.org/https://doi.org/10.1002/aisy.202400304

[48] Shao, H., Wang, L., Chen, R., Li, H., & Liu, Y. (2023). Safety-enhanced autonomous driving using interpretable sensor fusion transformer. In K. Liu, D. Kulic, & J. Ichnowski (Eds.), *Proceedings of the 6th conference on robot learning* (Vol. 205, pp. 726–737). PMLR. Retrieved from https://proceedings.mlr.press/v205/shao23a.html

[49] Singh, C., Inala, J. P., Galley, M., Caruana, R., & Gao, J. (2024). Rethinking interpretability in the era of large language models. Retrieved from https://arxiv.org/abs/2402.01761

[50] Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K., & Cilar, L. (2020). Interpretability of machine learning-based prediction models in healthcare. *WIREs Data Mining and Knowledge Discovery*, *10*(5), e1379. https://doi.org/https://doi.org/10.1002/widm.1379

[51] Sundararajan, M., & Najmi, A. (2020). The many shapley values for model explanation. In H. D. III & A. Singh (Eds.), *Proceedings of the 37th international conference on machine learning* (Vol. 119, pp. 9269–9278). PMLR. Retrieved from https://proceedings.mlr.press/v119/sundararajan20b.html

[52] Turpin, M., Michael, J., Perez, E., & Bowman, S. (2023). Language models dont always say what they think: Unfaithful explanations in chain-of-thought prompting. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems* (Vol. 36, pp. 74952–74965). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2023/file/ed3fea9033a80fea1376299fa7863f4a-Paper-Conference.pdf

[53] Wong, F., Zheng, E. J., Valeri, J. A., Donghia, N. M., Anahtar, M. N., Omori, S., Li, A., Cubillos-Ruiz, A., Krishnan, A., Jin, W., Manson, A. L., Friedrichs, J., Helbig, R., Hajian, B., Fiejtek, D. K., Wagner, F. F., Soutter, H. H., Earl, A. M., Stokes, J. M., Renner, L. D., & Collins, J. J. (2024). Discovery of a structural class of antibiotics with explainable deep learning. *Nature*, *626*(7997), 177–185. https://doi.org/10.1038/s41586-023-06887-8

[54] Yang, M., & Kim, B. (2019). Benchmarking attribution methods with relative feature importance. Retrieved from https://arxiv.org/abs/1907.09701

[55] Yang, Y., Lv, H., & Chen, N. (2023). A survey on ensemble learning under the era of deep learning. *Artificial Intelligence Review*, *56*(6), 5545–5589. https://doi.org/https://doi.org/10.1007/s10462-022-10283-5

[56] Zablocki, É., Ben-Younes, H., Pérez, P., & Cord, M. (2022). Explainability of deep vision-based autonomous driving systems: Review and challenges. *International Journal of Computer Vision*, *130*(10), 2425–2452. https://doi.org/10.1007/s11263-022-01657-x

[57] Zhang, Q., & Zhu, S.-C. (2018). Visual interpretability for deep learning: A survey. *Frontiers of Information Technology & Electronic Engineering*, *19*(1), 27–39. https://doi.org/https://doi.org/10.1631/FITEE.1700808