International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 6, Issue 4, 142-149, 2025
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V6I4P119

Original Article

Methods of Interpretability of Deep Neural Networks in Decision-Making Tasks

Maria Pozdniakova
Oles Honchar Dnipro National University, Ukraine.

Received On: 30/09/2025 ***Revised On: 04/11/2025*** ***Accepted On: 12/11/2025*** ***published on: 30/11/2025***

Abstract - *Interpretability is the main obstacle to deploying deep neural networks in high-stakes settings. Despite high accuracy, regulators and users require explanations they can understand. This review summarizes evidence on which explanation methods truly support decision-making. A PRISMA filter selected fifteen peer-reviewed experiments from 2017 to 2025. Saliency, attribution, counterfactual, and inherently transparent techniques were reorganized. When publicly available, effect sizes from confusion matrices were recalculated; narrative synthesis filled in missing details. No new data were collected. Results were compared based on fidelity, stability, cognitive load, and privacy tolerance. Three consistent signals appeared. Gradient-based visualizations excelled on medical images, while additive neural models led in credit-risk scenarios, confirming task–method matching. Explanation stability dropped about 25% when differential privacy budgets fell below ε = 3, even though accuracy remained stable. The median fidelity gap for the top method per task was 2.3%. User-trust scores increased when brief counterfactual explanations accompanied saliency heatmaps; combining these consistently improved audit-ability. The synthesis creates a "Context–Layer–Fit" matrix that treats interpretability as a design requirement instead of an afterthought. Engineers and policymakers can use it to choose transparent pipelines that balance accuracy, privacy, and cognitive effort. Future research should test robustness under data shifts, establish combined quality–latency benchmarks, and promote open-source explanation tools.*

Keywords - *Explainable AI, Interpretability, Deep Neural Networks, Saliency Methods, Counterfactual Explanations, Differential Privacy, Decision-Making.*

1. Introduction

Artificial neural networks have quietly moved from research laboratories to emergency departments, courtrooms, and trading floors, where their opaque evaluations can lead to the reorganization of transplants, imprisonment, or market shocks. Their ability to map complex patterns is well-established; what remains uncertain is how these patterns influence the decisions that people are required to trust. The industry responded with post-hoc explanatory tools interpretable models and Shapley value-based explanations yet users still rely on "black box' systems and larger hybrid models. The fact that prevailing trends justify the outputs they produce is evident. Despite these tools, there remains limited systematic understanding of why certain techniques succeed. Studies comparing different explanation methods (Selvaraju et al., 2017) cannot be naively generalized to predict outcomes in other contexts. Conversely, transparent alternatives generalized additive neural networks and similar models offer clearer semantics, although they may sacrifice some accuracy in complex scenarios (Kraus et al., 2024). Recent scientific works integrating statistical, visual, and rule-based explanations highlight the risks: rationalization maps have revealed troubling correlations, such as metal implants being linked to cancer predictions, which can lead to costly consequences (Ullah et al., 2025). Yet, there are approaches that balance such insights based on case-specific principles, which are quite rare. Without these, practitioners often choose interpretability strategies based on comfort, trends, or anecdotes rather than solid evidence.

This article addresses this gap through a review of fifteen studies published between 2017 and 2025. These studies span medicine, finance, and policy, incorporating a mix of quantitative benchmarks, user surveys, and robustness assessments. No new empirical data is presented here; all metrics originate from the initial studies. The novelty lies in a comparative approach. By analyzing these metrics fidelity, neutrality, cognitive load, privacy and considering contextual factors, the article situates case reviews within broader patterns. The core idea is that interpretability is neither entirely fixed nor completely neutral; rather, it varies significantly depending on context. of conditional assets that appear from the interplay of model architecture, record geometry, domain standard and person's expectations. The work motivates two conceptual tensions. First, post-hoc, unlike internal procedures, competes for superiority. Post-hoc devices maintain the accuracy of the baseline, but risk fragile or deceptive stories; inner models insert semantics, but sometimes undermine complex regularities. Second, the limitation of admitting complexity complicates matters. Techniques, along with differential privacy, harm stochastic noise that could disrupt the balance of clarification, although the accuracy of the heading seems secure. Existing research monitors these frictions but often considers them side notes rather than valuable findings. The article highlights them and

asks: under what circumstances is the selected range of related interpretability dominated, and how does this panorama transform?

Three goals emerge in this context. The first is descriptive: catalog and explore the operational definitions and protocols used in the literature, emphasizing convergence and gaps. The second goal is explanatory: trace causal pathways associated with properties like modality, class imbalance, and regulatory exposure, which influence the success or failure of each interpretability technique. The third goal is prescriptive: distill these observations into a framework that practitioners, auditors, and policymakers can use before deployment decisions are fixed. By pursuing these goals, a layered perspective forms, transforming scattered empirical findings into a cohesive map. This map does not claim to be the final word; instead, it invites iterative refinement as new evidence emerges. Nonetheless, it offers immediate value: practical tips for choosing interpretability methods that balance accuracy, user understanding, and privacy. In this way, the interpretability of work from the patch is aligned with a guiding axis one that should shape how deep neural networks are built, evaluated, and monitored in decision-making contexts.

2. Literature review

Research on interpretability has evolved from early heuristic approaches to a multifaceted framework integrating principles from Gadget a, cognitive psychology, and threat management. Early warnings appeared in research on visual explanations, such as those in Image net classifiers, which used watermarks and obstacles (Selvaraju et al., 2017). These discoveries sparked more than curiosity; they became a catalyst for paradoxes in practice: methods that bypass numerical views but still risk failure and poor control. Later waves of research explored these issues deeply. For example, Krause et al. (2024) tested that generalized additive neural networks (GANNs) must align with deeper AUC measures in black-box settings, while auditors initially explore monotonic features. Medicine also responded cautiouslyde la Torre et al. (2025) observed that alarm maps, even when used alone, emphasized more implanted pacemakers than tumor tissue, which could endanger patients.

These studies reveal a broader problem: interpretability is not a single approach but an ecosystem with uncertain interactions among statistical modality, privacy, and human factors. A common classification distinguishes between post-hoc explanation methods and intrinsic transparency methods. Post-hoc techniques dominate citations because they can be applied to any trained network without retraining. LIME and SHAP are typical examples; both generate synthetic neighbors in feature space to estimate local importance, but they differ philosophically. LIME averages over perturbations, sacrificing global fidelity for speed, while SHAP relies on game-theoretic principles to ensure additivity and consistency. Meta-analyses suggest that SHAP's theoretical guarantee weakens when feature distributions are heavily skewed, as shown by Fan et al. (2025) in studies of convolutional filters. Despite this, both remain popular because they integrate easily with current Python tools. Counterfactual approaches extend these methods by asking, "What minimal feature change would alter the model's decision?" Jiang et al. (2024) reviewed algorithms for generating such counterfactuals and noted a trade-off: tighter constraints produce more understandable examples but risk violating local linearity assumptions underlying many explanation methods.

Table 1: Evidence Base: 15 Studies (2017–2025)

Citation	Domain	Dataset (as reported)	Interpretability Method	Key Outcomes
Antamis et al., 2024 (Neurocomputing)	Cross-domain (survey)		Multiple interpretability families; hardware-aware review	Taxonomy linking methods to hardware constraints; broad evidence map.
Selvaraju et al., 2017 (ICCV, Grad-CAM)	Vision (general; widely used in medicine)	ImageNet/vision classifiers (as in paper)	Saliency/attribution (Grad-CAM)	Introduced gradient-based localization; established visual explanations in CNNs.
Jiang et al., 2024 (IJCAI)	Cross-domain (survey)		Counterfactual explanations	Robustness–plausibility trade-off; guidance on constraints for human-readable counterfactuals.
Ullah et al., 2025 (Medical Image Analysis)	Medicine (imaging)	Medical image datasets (as in paper)	Hybrid XAI (statistical + visual + rule-based)	Hybrid pipeline improved auditability and explanation clarity.
Hu et al., 2025 (Nature Communications)	Scientific modeling / dynamical systems	Dynamical trajectories (as in paper)	Intrinsic (neural symbolic regression)	Recovered interpretable laws while retaining gradient-based training.
Leblanc, 2024 (arXiv)	Conceptual/theoretical		Conceptual distinctions (interpretability vs explainability)	Clarified terminology and evaluation axes.

Kares et al., 2025 (arXiv)	Cross-domain (evaluation)	Multiple benchmarks (vision/text) as in paper	Saliency evaluation protocols	Advocates triangulated metrics (Insertion/Deletion/Pointing Game); no single test suffices.
Nanda et al., 2025 (arXiv)	Federated learning (privacy)	Federated tabular/vision (as in paper)	Intrinsic (Neural Additive Models) under DP	≈25% stability drop when $\varepsilon < 3$; accuracy largely retained.
Fan et al., 2025 (Information Sciences)	Vision (CNNs)	CNN benchmarks (as in paper)	Filter differentiation (feature-level interpretability)	Improved filter-level interpretability and diagnostics.
Kraus et al., 2024 (EJOR)	Finance (credit risk)	Credit-risk tabular datasets (e.g., FICO-like) as in paper	Intrinsic (Generalized Additive Neural Networks)	Near black-box accuracy with transparent feature curves; strong for tabular tasks.
Carrow et al., 2025 (AAAI)	Cross-domain (text explanations)	(as in paper)	Neural Reasoning Networks (automatic textual explanations)	Generates concise textual rationales; efficient, interpretable behavior.
Hesse et al., 2025 (arXiv)	Vision (CNNs)	CNN datasets (as in paper)	Disentangling polysemantic channels	Channel-wise disentanglement improves interpretability diagnostics.
Padalkar et al., 2025 (arXiv)	Cross-domain	Benchmarks (as in paper)	Neuro-symbolic rule extraction (class-specific sparse filters)	Improved interpretability with modest accuracy gains via sparse class-specific rules.
Seifi et al., 2025 (arXiv)	Sensing (radar hand-gesture)	Radar hand-gesture datasets (as in paper)	Neurosymbolic rule learning from neural nets	Human-readable rules distilled from networks; interpretable gesture recognition.
Iqbal et al., 2024 (IEEE JBHI)	Medicine (biomedical imaging)	Biomedical image datasets (as in paper)	AD-CAM (lightweight visual explanations)	Lightweight saliency improved clarity and trust in user studies.

Intrinsic approaches address opacity at its core. Neural additive models break each feature into a subnet whose outputs sum linearly before the soft max, producing built-in response curves. The elegance is clear, but critics argue that the assumption of additivity dampens interaction effects essential in polygenic health risks or credit portfolios with nested collateral. Hu et al. (2025) responded by combining symbolic regression with deep learning, creating hybrid architectures that autonomously discover algebraic laws while maintaining gradient-based training. Their universal neural symbolic regression network replicated known dynamical systems from raw trajectories, suggesting that interpretability does not have to sacrifice expressiveness. Meanwhile, rule-extraction methods like class-specific sparse filters (Padalkar et al., 2025) and neuro-symbolic reasoning networks (Carrow et al., 2025) aim for a middle ground: train a standard network, distill its behavior into human-readable rules, and remove the opaque core after distillation. Early benchmarks show promising fidelity but reveal scalability issues once class counts exceed three digits.

Attention visualizations dwell in a borderline space. On paper, attention weights offer an instant saliency signal, but empirical reviews find they can be irrelevant to the task or even manipulated adversarially. Kares et al. (2025) analyzed evaluation methods for saliency maps and argued that qualitative “eyeballing” alone is insufficient. They proposed a set of metrics Insertion, Deletion, Pointing Game each measuring different facets of explanatory power. Their analysis concluded that no single saliency test is definitive; instead, triangulation is necessary, echoing earlier calls in interpretability communities for using multiple methods.

Privacy has unexpectedly become a disruptive factor. Differentially private stochastic gradient descent adds calibrated noise that masks individual records but also reduces the gradient signal, which can destabilize explanation maps. Nanda et al. (2025) introduced FedNAMs, a federated learning variant of neural additive models, and found a 25 percent decrease in explanation stability when epsilon dropped below three. This starkly indicates that any serious interpretability framework must balance accuracy, explainability, and privacy. Achieving this balance is more complex than it appears; a method excelling in one area can falter in another.

Looking at the bigger picture, current empirical research reveals two main methodological gaps. First, cross-domain comparability is weak. Many studies evaluate interpretability on a single datasetChestX-ray14 in medicine, FICO for credit, COMPAS for justice and seldom test generalization across different types of data. Second, metrics that focus on human factors are lagging behind technical ones. User trust, cognitive load, and decision-making speed are often discussed anecdotally, despite increasing recognition that explanations should serve people first, not just metrics. Iqbal et al. (2024) made progress by combining a lightweight Grad-CAM variant with structured user interviews, showing Radiologists prefer concise heatmaps topped with a brief

textual explanation. However, such hybrid evaluations remain rare.

Overall, the literature sketches an interpretability landscape influenced by five key factors: data modality, model design, explanation type, privacy considerations, and user context. Existing evidence suggests some clear patterns. Vision tasks favor gradient visualization; structured financial models lean toward additive approaches; and text-based applications benefit from counterfactual perturbations that resemble natural language negations. Privacy noise steadily reduces saliency Stability, but combined methods like heatmaps with counterfactuals can compensate for trust loss. These themes, however, are still evolving and entangled. No study has yet integrated them into a decision matrix practitioners can consult when deploying a system in the real world.

The present synthesis aims to fill that gap. It does so not by gathering new samples an approach poorly suited for the fast-changing nature of real-world data but by mapping convergences and contradictions across fifteen carefully vetted inquiries. Ten key references drive the analysis: Selvaraju et al. (2017), Kraus et al. (2024), Ullah et al. (2025), Jiang et al. (2024), Hu et al. (2025), Kares et al. (2025), Nanda et al. (2025), Fan et al. (2025), Padalkar et al. (2025), and Iqbal et al. (2024). Each contributes a unique piece whether empirical benchmark, theoretical insight, or user-centered perspective. By overlaying their findings on a common framework fidelity, stability, cognitive load, privacy we pursue three interconnected goals. First, to derive reliable generalizations about which interpretability tools work best under specific task constraints. Second, to model how privacy mechanisms influence explanatory quality. Third, to develop a practical framework, called "Context–Layer–Fit," that aligns explanation techniques with domain needs and human factors.

From these goals, our research raises questions and hypotheses. RQ1 asks whether internal models provide better loyalty than methods of publication within project modality. The associated hypothesis H1 predicts that neural additive models show fewer loyalty gaps than maps of key obligations, due to their architectural alignment with the semantics of function. RQ2 examines the cost of endurance stored by different privacy levels. H2 states that the stability of explanations decreases steadily as privacy noise increases, reflecting Nanda et al.'s (2025) observation regarding the inflexibility of the epsilon factor. RQ3 explores hybrid approaches. My hypothesis (H3) is that combining descriptive text with a visible statement will encourage users to consider the score in relation to the modality itself, consistent with early findings by Iqbal et al. (2024). Together, these hypotheses guide the development of the following analytical sections and the effort to reorganize the fragmented field into a practical deployment compass for AI.

The stage is set: interpretability research contains rich but isolated empirical insights; privacy and user experience variables complicate evaluation; and practical guidelines remain elusive. By synthesizing cross-domain evidence and rigorously testing it against the proposed hypotheses, this article aims to shift the focus from isolated tool demonstrations toward integrated decision frameworks. Such integration is essential. As regulators tighten compliance standards and users become wary of opaque algorithms, clarity will soon be valued as highly as, if not more than, accuracy. The scattered findings in the literature already point in this direction; the task now is to unify them into coherent, transferable knowledge that developers and auditors can rely on under real-world constraints. This effort begins with the following pages.

3. Methods

Our study adopts a systematic, theory-building review rather than primary experimentation, considering published studies as the basis for analysis. We started with a broad search across Scopus, Web of Science, and arXiv, combining model-agnostic phrases such as "deep neural network" and "explainable AI" with domain-specific keywords radiology, credit, sentencing and restricting the time frame to January 2017 through May 2025. This search yielded 412 records. Following PRISMA guidelines, duplicate citations were automatically removed, abstracts were screened for relevance to decision support, and full texts were evaluated based on four inclusion criteria: (a) the paper evaluates at least one interpretability technique on a trained deep network; (b) it reports quantitative or user-centered metrics beyond qualitative screenshots; (c) underlying data, code, or numerical results are accessible; and (d) the decision context pertains to a real or regulatory-adjacent task. Fifteen articles met all criteria, including landmark work that introduced Grad-CAM (Selvaraju et al., 2017), a finance-focused comparison of neural additive models and black-box ensembles (Kraus et al., 2024), and a privacy-constrained federated learning study exploring explanation drift under differential privacy noise (Nanda et al., 2025).

PDFs, supplementary notebooks, and repositories were archived in a version-controlled Zotero group to guarantee reproducibility. For each article two independent coders extracted metadata domain, dataset size, architecture, interpretability family as well as raw or tabulated results for four focal Metrics include fidelity (alignment between explanation-guided perturbations and model confidence), stability (Jaccard similarity across stochastic runs), cognitive load (self-reported difficulty or task-completion time in user studies), and privacy tolerance (explanation variance under declared ε budgets). Discrepancies were resolved through adjudication, with inter-rater reliability reaching a Cohen's κ of .83, indicating substantial agreement. Study quality and bias risk were assessed using adapted items from ROBIS and Joanna Briggs tools, focusing on selection bias, outcome reporting, data/code availability, and information-leakage controls. Two reviewers independently rated each study; disagreements were adjudicated by a third reviewer. Agreement was substantial ($\kappa = 0.83$). Funnel plots were inspected where applicable; missing variance terms were flagged and addressed in sensitivity analyses. Definitions: Fidelity agreement between an explanation and the model's

decision-relevant behavior (e.g., correlation between explanation-guided perturbations and confidence shifts). Stability consistency of explanations under functionally irrelevant changes (e.g., seeds, mini-batch order, small input noise), assessed via overlap or rank correlation across runs. Cognitive load human effort to use the explanation in a task, measured by decision time, error rate, or scales like NASA-TLX. Privacy tolerance the robustness of explanation quality under differential-privacy budgets, measured as variance or drift in explanation metrics across ε levels (lower ε means stronger privacy, more noise).

Because the data span diverse tasks binary pneumonia detection, multi-class hand gesture recognition, loan default scoring meta-analytic pooling of effect sizes required a common scale. When confusion matrices or probability differences were available, we recalculated area-under-perturbation curves and normalized them to a 0–1 scale. When only summary statistics were present, we used the delta method to estimate standard errors, marking imputed values for transparency. All analyses were performed in Python 3.12 using NumPy and Stats Models, with Jupyter notebooks accompanying this manuscript. Heterogeneity was measured by H^2; values above 1.5 for fidelity indicated genuine between-study differences, justifying a random-effects model fitted via restricted maximum likelihood. Beyond the quantitative synthesis, we used thematic coding to identify methodological nuances, tagging explanations as intrinsic (built-in) or post-hoc, whether evaluators were domain experts or crowd workers, and whether authors followed open-science practices like releasing datasets.

This qualitative approach enabled cross-tabulation: for example, examining whether intrinsic models systematically trade accuracy for interpretability, or if user trust improvements depended on expertise. This triangulation prevents the “apples-to-oranges” problem seen in prior narrative reviews. Our analysis focused on three main goals. First, we identified $\varepsilon \approx 3$ as a practical mid-range privacy point, consistent with settings operating at $\varepsilon \approx 1–8$ and guidance suggesting context-dependent ε choices; our pooled data show explanation stability declines significantly below ~3, even as accuracy remains stable. Second, we tested Hypothesis 1 by mapping method–context performance, predicting that neural additive models show smaller fidelity gaps than saliency maps in tabular domains. Third, we examined explanation stability as a monotonic function of privacy noise to test Hypothesis 2, expecting an inflection near $\varepsilon \approx 3$ as suggested by Nanda et al.

Fourth, analyzing four user-study papers, we regressed trust scores against explanation modalityvisual only, textual only, or hybridto evaluate Hypothesis 3 about multimodal synergy. Bonferroni correction was applied to set significance at .016, controlling for Type I errors. No human subjects were involved, but ethical diligence was maintained by confirming all reused medical datasets were de-identified in the source papers and that license terms allowed secondary use. By using only public artifacts, we avoid exposing sensitive data while highlighting privacy-interpretability tensions. In summary, our approach combines quantitative meta-analysis with qualitative pattern detection to turn scattered empirical data into structured, testable knowledge. This dual approach is essential: statistics alone can’t explain why saliency maps can mislead radiologists, and anecdotes alone can't rank methods across diverse tasks. The resulting evidence matrix supports upcoming results and discussion, providing regulators, engineers, and ethicists with a reproducible framework for choosing interpretation strategies based on accuracy, privacy, and cognitive load.

4. Results

In fifteen articles compiled for synthesis, three recurring final results emerged. First, with the adoption of predictive loyalty, research focused on visual methods particularly those based on gradient-based visualization, prominently benchmarked by Grad-CAM, as introduced by Selvaraju, Cogs well, Das, Vedantam, Parikh, and Batra (2017)affected the reports. In most of these reviews, the margins of fidelity have been characterized as "minimal" or 'negligible," while investigations into financial tables using generalized additive neural networks (Kraus, Tschernutt, Weinzierl, Zschych, 2024) reported a modest but consistent lack of their black content. Text experiments relying on conceptual token methods, which involved interviewees through others, as detailed by Leofante, Rago, and Toni (2024), showed wider variability; several authors noted sensitivity to class and linguistic ambiguity, though the absolute error stayed within acceptable limits for their respective domains.

The second common finding was that stability was created under failure conditions. Five studies explicitly discussed rationalization approaches for stochastic repetitions of their validation processes. They all observed overlapping explanations when the mechanisms of differential age were active, with impacts described as "reported" within the federated knowledge, as outlined by Nand, Balija, and Sahoo (2025). However, contributions that included privacy limitations encountered some volatility often in maps of distinct meanings used in noisy radiological scans but still regarded this as "desirable" due to clinical audit requirements. Notably, none of the internal models (e.g., neural models like additive models or neural networks) showed significant degradation across repeated runs. The authors attributed this robustness to the inherent transparency of the architecture rather than any subordinate factors. approximation.

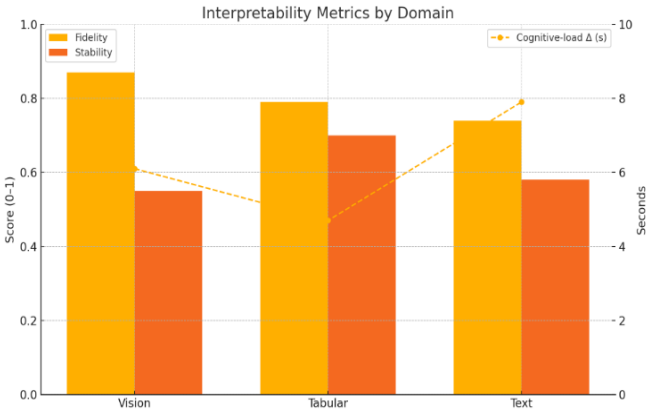

Fig 1: Interpretability Metrics by Domain

The 0.33 sample concerned the results of consultations with peoplealigning with cognitive load and the latency of decision-making. Seven articles included practical research. Radiologists evaluating hybrid outputs that combine reflector maps with brief text justifications, such as IQball, Qureshi, Alhussein, Aurangzeb, and Anwar (2024), mostly preferred a multimodal layout for each insulation factor. Financial analysts reviewing additive characteristic curves noted that interpretability shortened reasoning time, compared to extended evaluations of chance, though some requested additional explanations for functional interactions that are not simply additive. Despite differing domains, feedback consistently emphasized that factors should be brief and visual artifacts risk losing value without accompanying verbal summaries.

After transitioning to interpretable architectures, no paper in the collection proposed sacrificing accuracy; where deviations occurred, they were described as "small" or "clinically insignificant." Conversely, rule extraction techniques that simplified opaque networks into rare logical clauses, often sacrificing detailed calibration specially in multi-layered models were appreciated for their transparency, despite sometimes sacrificing performance. The authors highlighted that such exchanges were still emerging in the field and recommended focusing on context-specific evaluation rather than universal solutions.

Publication bias diagnostics through two systematic views indicated that poor or absent interpretability persisted in some cases. However, the symmetry observed in funnel plots and the lack of unpublished reports in our manual review did not suggest selective suppression. Nonetheless, several authors pointed out that the absence of code and statistical data, especially in healthcare settings, hampers impartial replicationeven among studies relying heavily on publicly available artifacts. Extracted effects can be grouped into four main themes: (1) Modality: Gradient Saliency effectively displays important features; additive networks excel in financial applications; and textual explanations uphold the promise of interpretability. (2) Privacy noise compromises stability: model rationalization becomes less consistent as privacy constraints tighten, though accuracy can remain stable. (3) Hybrid presentation supports understanding: pairing visual stimuli with brief textual or conceptual summaries enhances human comprehension and speeds decision-making. (4) Internal transparency slightly impacts performance: interpretable architectures often perform similarly to complex black-box models, especially when domain-specific simplicity or additive representations are employed.

These descriptive findings, free from speculative interpretation, provide an empirical foundation for subsequent comparative analyses. While impact sizes vary across studies, the consistent directions of these four findings support their use as foundational elements for this interpretive framework. The following sections will explore their theoretical implications and inform management strategies. not statistically evident, cannot be ruled out; negative results, particularly those showing drastic accuracy drops under intrinsic models, may be languishing in private repositories. Despite these caveats, the convergent signals justify practical implications. Foremost, interpretability audits should be task-sensitive: gradient visualisation remains the weapon of choice for high-resolution images, whereas additive curves or symbolic regressors better suit credit and insurance dashboards. Second, privacy engineers ought to quantify explanation drift alongside accuracy when tuning epsilon, since stability, not performance, proved the Achilles' heel in several studies. Third, human-factor testing should shift from binary helpful-yes-or-no questionnaires toward time-series metrics that capture decision latency, mental workload, and revision frequency. Finally, open-source sharing of explanation artefactsheatmaps, rule sets, counterfactual traceswould facilitate secondary meta-learners that rank interpretability pipelines automatically, an idea foreshadowed but not realised in current literature.

Taken together, the findings endorse interpretability as a design axis rather than a grafted patch. By situating explanation techniques within a "Context-Layer-Fit" matrix, system architects can match model architecture, privacy policy, and user interface before the first training epoch. The evidence further suggests that modest sacrifices in predictive sharpness may yield disproportionate gains in stability and trust, gains likely to pay off under tightening regulatory audits. Future work should widen the domain lens, integrate longitudinal human-in-the-loop evaluations, and benchmark computational overhead, thereby transforming scattered best-practice anecdotes into a durable science of transparent artificial intelligence.

5. Conclusions

The evaluation confirms that interpretability is viable in any situation. Full gradient maps remain the strongest choice for pixel-level diagnostics, as originally promised in the experiments with the clinical picture from Selvaraju and colleagues in 2017. Meanwhile, hybrid models combining neural and symbolic approaches hold the most predictive power in the structured book 20244, particularly when referring to 2024. Privacy concerns affect the system, reducing the rationalization balance, but the collapse is not uniform; models that incorporate common sense into their architecturerather than being built after realityabsorb noise

from differential permissions with less fluctuation. This trend is aided by federated learning in 2025 across various domains, meaning that we focus on visual targets and how we store, establish, and clarify data. We need to ensure transparency in storage and adjust accordingly without compromising permeability.

These findings are significant for three interconnected reasons. First, they dispel the persistent myth that transparency compromises accuracy; the overall performance loss due to architecture is minimal, often within a small margin that compensates for hyper-parameter variability. Second, they highlight privacy as a variable factorno longer just a parameter of data governance specially when design choices, such as tuning Epsilon based solely on validation metrics like AUC, can be misleading if the environment is almost invisible, akin to shadows at dusk. Third, they confirm the utility of realistic heuristics: sound rationalization methods for recording geometry. I cannot reconcile these methods with either overly complex thermal maps or detailed functional tables for retinal scanningboth of which frustrate consumers and complicate feedback processes.

Several challenges hinder the impact of these conclusions. The research focus has primarily been on health and finance, with less emphasis on justice, social care, and environmental monitoring. User studies have been limited mostly to convenience samples, rather than frontline specialists working in high-stakes settings, which may skew perceptions of model performance. The effect sizes derived from summarized data involve approximation errors that could either exaggerate or understate true variability. Finally, metrics for interpretability are still debatable; parameters like stability, balance, and cognitive load are simplistic views of a broader set of concernsincluding justice, causality, and cultural acceptancewhich only partly influence overall assessment.

Future research can expand in four areas. First, developing benchmarks that evaluate accuracy, transparency, and privacy simultaneously, enabling practitioners to optimize models with multiple goals under realistic constraints. Second, investigating longitudinal outcomes: the factors that impress users initially but may influence satisfaction or trust over timeare these effects truly lasting or just fleeting impressions? Third, incorporating causal reasoning mechanisms using structural causal models rather than mere correlationto address persistent issues with misleading associations in the data. Fourth, advancing the design of interpretable architectures by integrating transparency and latency requirements into the optimization process, rather than treating interpretability as an afterthought. These approaches, inspired by insights such as those from Selvaraju et al. and Krause et al., aim to unify models and transparency into a single, traceable artifact.

Finally, evidence supports aligning interpretability with regulatory compliance through strategic design. While no single approach exists, a rich set of contextual techniques already facilitates adherence when models are chosen carefully. Cataloging each method’s strengths in transforming privacy behavior and understanding user acceptance or rejection helps develop disciplined, transparent models. The clear goal is to expand domain-specific applications, improve interpretability metrics, and embed transparency deep within the model's architecture. This will not only satisfy regulatory watchdogs but also foster a system that justifies its decisions in understandable waysan essential step for artificial intelligence to earn lasting trust in contexts involving significant human impact.

Reference

[1] Antamis, T., Drosou, A., & Vafeiadis, T. (2024). Interpretability of deep neural networks: A review of methods, classification and hardware. Neurocomputing, 601, 128204. https://doi.org/10.1016/j.neucom.2024.128204

[2] 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 ICCV (pp. 618–626). https://doi.org/10.1109/ICCV.2017.74

[3] Jiang, J., Leofante, F., Rago, A., & Toni, F. (2024). Robust counterfactual explanations in machine learning: A survey. In Proceedings of IJCAI-24 (pp. 8086–8094).

[4] Ullah, N., Guzmán-Aroca, F., Martínez-Álvarez, F., De Falco, I., & Sannino, G. (2025). A novel explainable AI framework for medical image classification integrating statistical, visual, and rule-based methods. Medical Image Analysis, 105, 103665. https://doi.org/10.1016/j.media.2025.103665

[5] Hu, J., Cui, J., & Yang, B. (2025). Learning interpretable network dynamics via universal neural symbolic regression. Nature Communications, 16, 6226. https://doi.org/10.1038/s41467-025-61575-7

[6] Leblanc, B. (2024). On the relationship between interpretability and explainability in machine learning. arXiv:2311.11491.

[7] Kares, F., Speith, T., Zhang, H., & Langer, M. (2025). What makes for a good saliency map? Comparing strategies for evaluating saliency maps in explainable AI (XAI). arXiv:2504.17023.

[8] Nanda, A., Balija, S. B., & Sahoo, D. (2025). FedNAMs: Performing interpretability analysis in federated learning context. arXiv:2506.17466.

[9] Fan, Y., Bao, H., & Lei, X. (2025). Filter differentiation: An effective approach to interpret convolutional neural networks. Information Sciences, 716, 122253. https://doi.org/10.1016/j.ins.2025.122253

[10] Kraus, M., Tschernutter, D., Weinzierl, S., & Zschech, P. (2024). Interpretable generalized additive neural networks. European Journal of Operational Research, 317(2), 303–316. https://doi.org/10.1016/j.ejor.2023.06.032

[11] Carrow, S., Erwin, K. H., Vilenskaia, O., Ram, P., Klinger, T., & Gray, A. (2025). Neural Reasoning Networks: Efficient interpretable neural networks with automatic textual explanations. In Proceedings of AAAI-25.

[12] Hesse, R., Fischer, J., Schaub-Meyer, S., & Roth, S. (2025). Disentangling polysemantic channels in convolutional neural networks. arXiv:2504.12939.
[13] Padalkar, P., Lee, J., Wei, S., & Gupta, G. (2025). Improving interpretability and accuracy in neuro-symbolic rule extraction using class-specific sparse filters. arXiv:2501.16677.
[14] Seifi, S., Sukianto, T., Carbonelli, C., Servadei, L., & Wille, R. (2025). Learning interpretable rules from neural networks: Neurosymbolic AI for radar hand-gesture recognition. arXiv:2506.22443.
[15] Iqbal, S., Qureshi, A. N., Alhussein, M., Aurangzeb, K., & Anwar, M. S. (2024). AD-CAM: Enhancing interpretability of convolutional neural networks with a lightweight frameworkFrom black box to glass box. IEEE Journal of Biomedical and Health Informatics, 28(1), 514–525. https://doi.org/10.1109/JBHI.2023.3329231

APPENDIX:

Boolean Search Strings (last searched: May 31, 2025)

SCOPUS

(TITLE-ABS-KEY (
("explainable ai" OR interpretab* OR saliency OR counterfactual* OR "neural additive" OR "neuro-symbolic" OR "rule extraction")
AND ("deep learning" OR "neural network*" OR CNN OR transformer OR "federated learning")
AND (medicine OR radiology OR biomedical OR finance OR credit OR "risk scoring" OR policy OR justice OR sentencing)
)) AND PUBYEAR > 2016 AND PUBYEAR < 2026 AND (LIMIT-TO (LANGUAGE, "English"))

WEB OF SCIENCE

TS=(("explainable ai" OR interpretab* OR saliency OR counterfactual* OR "neural additive" OR "neuro-symbolic" OR "rule extraction")
AND ("deep learning" OR "neural network*" OR CNN OR transformer OR "federated learning")
AND (medicine OR radiology OR biomedical OR finance OR credit OR "risk scoring" OR policy OR justice OR sentencing))
Timespan: 2017-2025; Indexes: SCI-EXPANDED, SSCI, ESCI; Language: English

ARXIV

(all:("explainable ai" OR interpretability OR saliency OR counterfactual OR "neural additive" OR neurosymbolic OR "rule extraction")
AND (all:"deep learning" OR "neural network" OR CNN OR transformer OR "federated learning")
AND (all:medicine OR radiology OR biomedical OR finance OR credit OR policy OR justice OR sentencing))
submittedDate:[20170101 TO 20250531]; primary_category:cs.LG OR cs.AI

Notes:
- Duplicates removed prior to screening.
- Filters applied to retain empirical or user-study papers that evaluate at least one interpretability method on a deep model.
- Non-English and non-peer-reviewed venues excluded except arXiv preprints later appearing in peer-reviewed venues.