SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Explainable AI In Deep Learning: A Comparative Taxonomy And Evaluation Framework For Post-Hoc Interpretability Methods

Dr. Manisha Anil Kumbhar[1], Neha Shah[2], Dr Vidya Sunil Gavekar[3], Anagha Suresh Chaudhari[4], Megha Mane[5], Amar Anant Shinde[6], Dr. Prashant Bhavarlal Chordiya[7]

[1]Professor & Director, Department of MCA, Suryadatta Institute of Business Management & Technology, Pune, India, manishakumbhar05@gmail.com 0000-0001-8763-5372
[2]Assistant Professor, Department of MCA, Suryadatta Institute of Business Management & Technology, Pune, India, Neha.shah@suryadatta.edu.in, 0009-0004-4488-8360
[3]Professor, Department of MCA, Suryadatta Institute of Management and Mass Communication, Pune, India, vidyagavekar@gmail.com, 0000-0001-8422-560X
[4]Assistant Professor, Department of MCA, Suryadatta Institute of Business Management & Technology, Pune, India, anaghachaudhari.info@gmail.com, 0009-0000-4058-9575
[5]Assistant Professor, Department MCA, Suryadatta Institute of Business Management and Technology, Pune, India, meghamane84@gmail.com, 0009-0009-3360-7056
[6]Assistant Professor, Department of MCA, Suryadatta Institute of Management & Mass Communication, Pune, India, amarslives@gmail.com, ORCID - 0009-0004-6318-7287
[7]Associate Professor, D.Y. Patil Institute of Master of Computer Applications and Management, Pune, India, prashant.chordiya@gmail.com, 0009-0005-6227-2583

Abstract

The issue of opacity in the decision making process is getting more relevant nowadays with the rapid deployment of deep learning models in the domains like healthcare, cybersecurity, and education. Explaining the decisions made by artificial intelligence models becomes more challenging with the increasing complexity of the tasks that need to be solved. Explainable Artificial Intelligence (XAI) emerged as a multidisciplinary research field that aims to create human-friendly explanations for predictions of sophisticated models without substantial loss in prediction performance. This paper presents an integrative literature review and a comparative framework for post-hoc XAI methods for deep learning. Using the taxonomy of XAI approaches that is based on differentiation into intrinsic and post-hoc explanations, local and global explanations, and model-specific and model-agnostic explanations, the paper highlights three main groups of methods: perturbation-based methods (e.g., LIME), axiomatic game-theoretic methods (e.g., SHAP), and gradient/attribution-based methods (e.g., Saliency Maps, Grad-CAM, Integrated Gradients, DeepLIFT, and Layer-wise Relevance Propagation). Comparative taxonomy chart, explanation pipeline for the three families of methods, and multi-dimensional qualitative comparison of selected methods are provided as a way of visualization of trade-offs regarding model-agnosticism, computational efficiency, theoretical background, stability of the explanation, and human-comprehensibility. Literature review also involves discussion of the applied domains of the methods, such as malware classification, medical imaging, and educational text classification. Open challenges in the area of XAI research are presented as follows: lack of a commonly agreed set of evaluation metrics, sensitivity of the explanations to small changes in the input data, and trade-off between fidelity and interpretability of the explanations.

Keywords: explainable AI; interpretability; deep learning; SHAP; LIME; Grad-CAM; feature attribution; model transparency

1. Introduction

State-of-the-art decision-making frameworks in fields ranging from medicine, to loans approvals, to cybersecurity utilize neural architectures of deep learning for outperforming humans. The problem is that the majority of decision-making models, such as deep convolutional and recurrent neural networks, transformers, or ensembles of those do not

provide opportunities to interpret the decision-making process. As a result, although a model has high predictive abilities, the end user does not have any information about the rationale behind the particular decision, which poses an issue when decision justification rather than mere precision is needed due to reasons of accountability, regulation, transparency, or medical safety. The field of explainable artificial intelligence (XAI) can be described as interdisciplinary science aimed at solving the interpretability problem with techniques that will make the behavior of black box understandable for humans.

The literature on XAI experienced fast development from several directions which can be described as follows: perturbation-based local explanations, gradient-based attribution techniques specific to differentiable architectures, and axiomatic explanations based on cooperative game theory. Practical applications of recent research efforts in this field cover a range of domains starting from malware images classification, breast cancer mammography, pediatric sleep apnea recognition via physiological signal analysis to educational text classification and much more. All applications involve usage of combinations of at least two different XAI methods, typically including SHAP, LIME, and Grad-CAM among other methods. Nevertheless, although there is a large number of case studies in XAI, there is a lack of comparative analysis and discussion of different methods within the same framework. This paper tries to fill this gap.

The contributions of this paper are threefold. First, the authors have provided a structured classification scheme for organizing the important aspects from the existing literature of XAI. This classification scheme distinguishes between the two types of explanations that could be produced by an algorithm, i.e., intrinsic or post-hoc; the type of scope of the explanation, i.e., local or global; and the type of explanation, i.e., model-agnostic or model-specific. Second, the authors have provided a comparative framework which will help in understanding the decisions and trade-offs that should be considered while comparing the various techniques used in XAI, such as LIME, SHAP, Grad-CAM, and Integrated Gradients using the three families of explanation concept.

2. Literature Review

2.1 Defining Interpretability and Explainability

A clear but somewhat implicit differentiation exists between the two terms in XAI literature. Interpretability deals with the level of comprehensibility of the reasons behind the model's decision. Explainability is more comprehensive in the sense that it describes the rationale behind the decision in human understandable language. The call by Doshi-Velez and Kim for the science of interpretable machine learning was based on the fact that the topic had never been clearly defined until their research. According to Arrieta et al., widely cited taxonomy paper frames XAI as pursuing "responsible AI" objectives — trustworthiness, causality, transferability, informativeness, fairness, and privacy awareness — that extend beyond raw predictive accuracy.

2.2 The Interpretability–Performance Trade-off

A recurring theme across the XAI literature is the presumed trade-off between model complexity (and the performance gains it often affords) and interpretability: simple, transparent models such as linear regression, logistic regression, and shallow decision trees are inherently interpretable but often less accurate on complex tasks, while deep neural networks achieve state-of-the-art performance at the cost of interpretability. This trade-off motivates two broad design strategies documented in the literature: intrinsic (or ante-hoc) interpretability, in which models are constructed to be interpretable by design, and post-hoc explainability, in which a separate explanation method is applied to an already-trained black-box model. Recent work has begun to question the strictness of this trade-off, proposing ante-hoc architectures such as the Deeply Explainable Artificial Neural Network (DxANN), which embeds per-sample, per-feature explanation generation directly into the forward pass rather than relying on external post-hoc methods.

2.3 Perturbation-Based Local Explanations: LIME

Ribeiro, Singh, and Guestrin's Local Interpretable Model-agnostic Explanations (LIME) method explains an individual prediction by perturbing the input around the instance of interest, observing the corresponding changes in model output, and fitting an interpretable surrogate model (typically sparse linear regression) that is locally faithful to the black-box model's behavior in that neighborhood. Because LIME treats the underlying model purely as a query-able function, it is fully model-agnostic and can be applied to tabular, text, or image data with an appropriate perturbation scheme (e.g., toggling superpixels on or off for images). LIME's principal limitations, documented extensively in follow-up work, include sensitivity to the choice of perturbation scheme and neighborhood size, and instability of the resulting explanations under small changes to the input or random sampling seed.

2.4 Game-Theoretic Explanations: SHAP

Lundberg and Lee's SHapley Additive exPlanations (SHAP) framework grounds feature attribution in Shapley values from cooperative game theory, originally developed by Shapley to fairly distribute payoff among players based on their average marginal contribution across all possible coalitions. Applied to model explanation, each input feature is treated as a "player," and its Shapley value quantifies its average marginal contribution to the model's prediction across all possible subsets (coalitions) of the other features. Lundberg and Lee showed that several existing attribution methods,

including LIME, can be understood as special cases within a unified additive feature-attribution framework, and that SHAP values satisfy desirable theoretical properties — local accuracy, missingness, and consistency — that are not guaranteed by all alternative methods. The principal practical limitation of exact SHAP computation is its exponential computational complexity in the number of features, which has motivated a substantial line of research into faster approximation algorithms.

2.5 Gradient- and Attribution-Based Explanations

A parallel family of methods exploits the differentiability of deep neural networks to compute feature attributions directly from gradients. Simonyan, Vedaldi, and Zisserman's early saliency-map approach visualizes the gradient of a class score with respect to input pixels, highlighting regions that most influence the prediction. Selvaraju et al.'s Gradient-weighted Class Activation Mapping (Grad-CAM) generates class-discriminative localization maps for convolutional neural networks by using gradients flowing into the final convolutional layer, producing coarse heat maps that highlight the spatial regions most responsible for a given prediction; a journal-length extension was later published in the International Journal of Computer Vision. Sundararajan, Taly, and Yan's Integrated Gradients method addresses the gradient-saturation problem — in which raw gradients become uninformative when a model's output function saturates — by integrating gradients along a straight-line path from a baseline (e.g., an all-black image) to the actual input, satisfying axioms of sensitivity and implementation invariance. Related attribution methods include DeepLIFT, which propagates activation differences relative to a reference input, and Layer-wise Relevance Propagation (LRP), which redistributes a model's output prediction backward through the network as “relevance” assigned to each input feature.

2.6 Evaluating the Quality of Explanations

Yet another line of research addresses the problem of measuring the quality of explanations. According to Guidotti et al., whose survey of the literature on the explanation techniques of black-box models is often cited, the approaches are classified based on the type of the explanator (decision tree, rule set, feature importance, and exemplar), and the type of the task to be done. Commonly proposed evaluation criteria include fidelity (how accurately the explanation reflects the underlying model's true reasoning), stability or robustness (whether similar inputs yield similar explanations), comprehensibility (whether human users can correctly interpret the explanation), and computational efficiency. Molnar's widely used reference text on interpretable machine learning further distinguishes between evaluating explanation methods analytically (via mathematical properties, as SHAP's axioms permit) and empirically (via human-subject studies of comprehension and trust).

Table 1. Core Dimensions Used to Classify XAI Methods

Dimension	Categories	Illustrative Examples
Timing of explanation	Intrinsic (ante-hoc) vs. Post-hoc	Decision trees (intrinsic); LIME, SHAP (post-hoc)
Scope of explanation	Local (single prediction) vs. Global (whole model)	LIME, Grad-CAM (local); permutation importance (global)
Model dependency	Model-agnostic vs. Model-specific	LIME, SHAP (agnostic); Grad-CAM, Integrated Gradients (specific to differentiable models)
Underlying mechanism	Perturbation-based, gradient-based, or axiomatic/game-theoretic	LIME (perturbation); Grad-CAM (gradient); SHAP (axiomatic)
Output form	Feature-importance scores, saliency/heat maps, or rule-based rationale	SHAP values; Grad-CAM heat maps; extracted decision rules

3. Comparative Framework

Building on the dimensions in Table 1, Figure 1 organizes the methods reviewed in Section 2 into a single taxonomy tree, and Figure 2 depicts the three principal methodological families — perturbation-based, gradient/attribution-based, and axiomatic/game-theoretic — as parallel pipelines that each transform a trained black-box model and an input instance into a human-interpretable explanation output

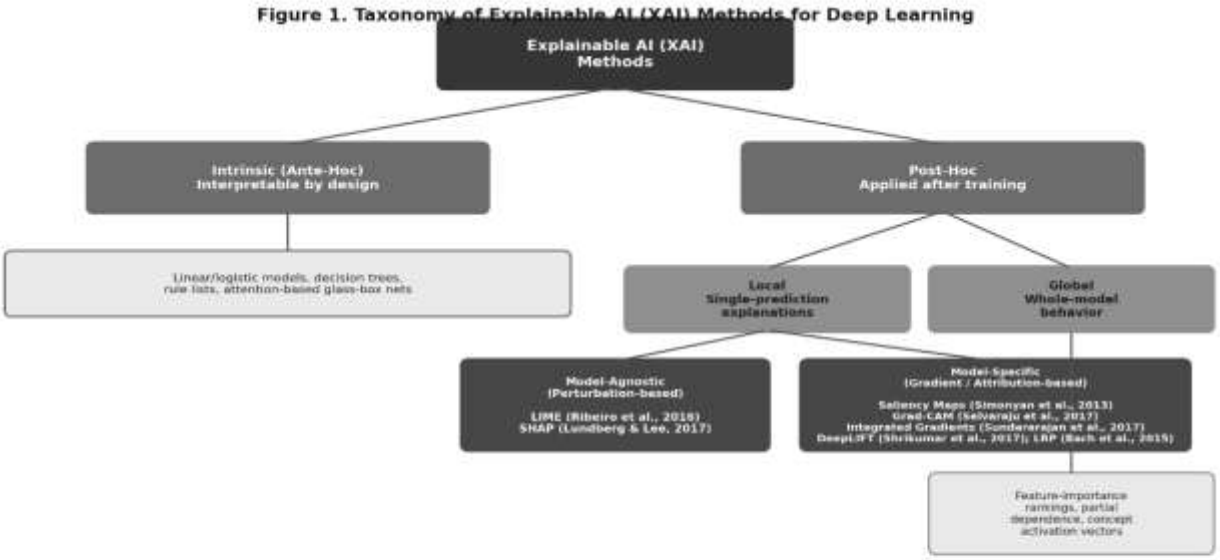

Figure 1. Taxonomy of explainable AI (XAI) methods for deep learning.

As Figure 1 illustrates, the intrinsic/post-hoc distinction operates at the top level of the taxonomy, separating models that are interpretable by construction from the much larger and more actively researched space of post-hoc methods applied to already-trained black-box models. Within the post-hoc branch, the local/global distinction determines whether an explanation addresses a single prediction or characterizes overall model behavior, while the model-agnostic/model-specific distinction determines whether the method can be applied to any predictive model or exploits properties specific to differentiable architectures.

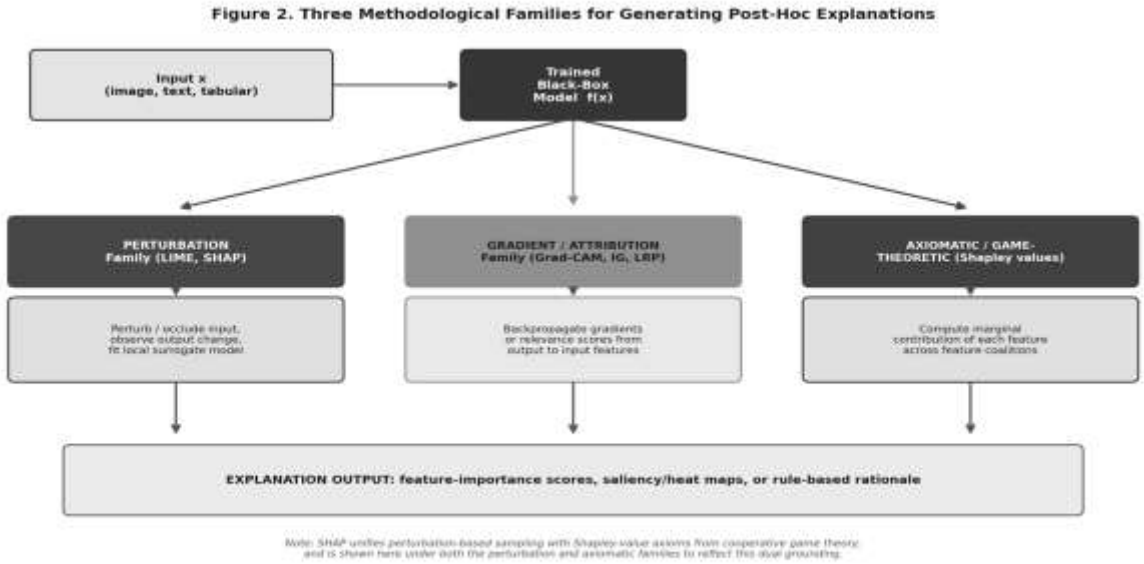

Figure 2. Three methodological families for generating post-hoc explanations.

Figure 2 highlights that, despite their differing theoretical foundations, all three families converge on structurally similar outputs — feature-importance scores, saliency or heat maps, or rule-based rationales — that are then interpreted by a human user. This convergence at the output stage is one reason multiple methods are frequently applied in combination in the applied literature (see Section 5), since agreement between methodologically distinct explanations can serve as an informal robustness check on the resulting interpretation.

3.1 Qualitative Comparison of Representative Methods

Since there are some factors that make these approaches different, Figure 3 represents a comparison between four approaches that are widely used, namely, LIME, SHAP, Grad-CAM, and Integrated Gradients, with respect to five aspects: model agnosticism, efficiency, theoretical underpinning, explanation stability, and human interpretability. It should be emphasized that the scores indicated on the diagram do not present the outcomes of any testing; rather, they reflect the properties that have been discussed in the literature.

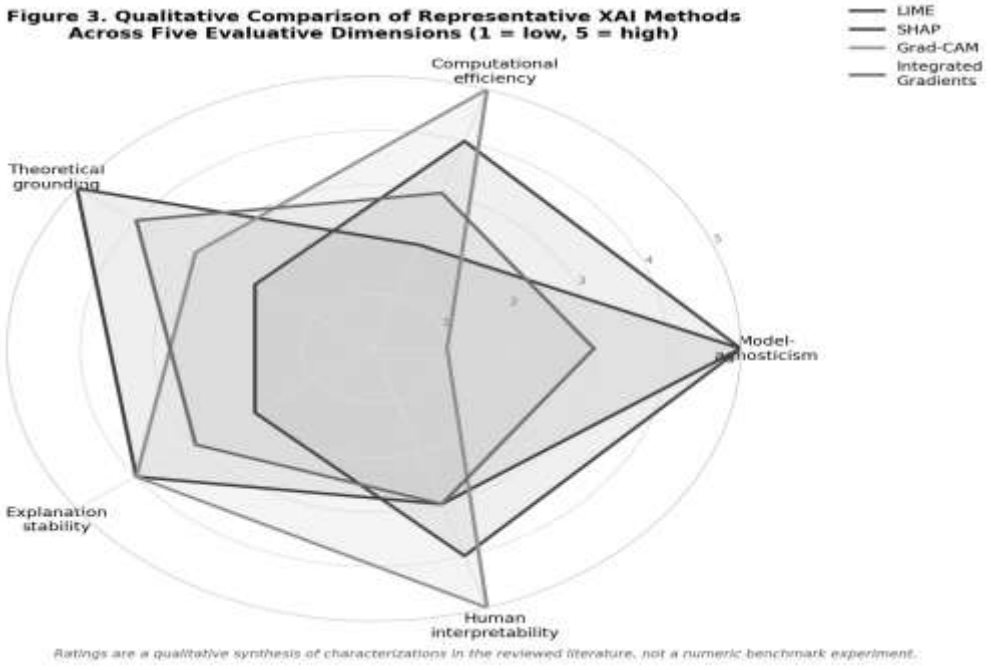

Figure 3. Qualitative comparison of representative XAI methods across five evaluative dimensions.

The patterns found in this synthesis are the following: The technique known as Grad-CAM is computationally efficient and can interpret images well, but this technique has limitations because it is designed specifically to work with convolutional neural networks. SHAP is considered the most theoretical technique of the four; SHAP is the only one that has been proven to have the property of uniqueness, but the problem with SHAP is that it is computationally expensive, and hence the approximations are normally applied. LIME technique possesses model agnosticism and computational efficiency, but according to the literature, it lacks stability. Finally, IG is somewhere between these two techniques.

Table 2. Comparative Summary of Major Post-Hoc XAI Methods

Method	Source	Family	Scope	Model Dependency
LIME	Ribeiro, Singh, & Guestrin (2016)	Perturbation-based	Local	Model-agnostic
SHAP	Lundberg & Lee (2017)	Axiomatic / game-theoretic	Local (aggregable to global)	Model-agnostic
Saliency Maps	Simonyan, Vedaldi, & Zisserman (2013)	Gradient-based	Local	Model-specific (differentiable)

Method	Source	Family	Scope	Model Dependency
Grad-CAM	Selvaraju et al. (2017)	Gradient-based	Local	Model-specific (CNNs)
Integrated Gradients	Sundararajan, Taly, & Yan (2017)	Gradient-based / axiomatic	Local	Model-specific (differentiable)
DeepLIFT	Shrikumar, Greenside, & Kundaje (2017)	Gradient/reference-based	Local	Model-specific (differentiable)
Layer-wise Relevance Propagation (LRP)	Bach et al. (2015)	Relevance-propagation	Local	Model-specific (differentiable)

4. Methodology

This research applies an integrative literature review methodology that is appropriate for the comprehensive review of the technologically diverse and rapidly developing area of research on XAI. The sources have been selected based on the following criteria: papers where LIME, SHAP, Grad-CAM, Integrated Gradients, DeepLIFT, and LRP were proposed; survey papers (Arrieta et al., 2020; Guidotti et al., 2018; Doshi-Velez & Kim, 2017); recent (2024-2026) application-oriented research in the domains of cybersecurity, medical images analysis, and education – these sources have been chosen in order to support the applied-domain review in Section 5. The sources have been synthesized by themes in the taxonomy (Section 3), comparative framework (Section 3.1), and applied-domain survey (Section 5) presented in this paper. As it is an integrative literature review paper, it does not offer experimental benchmarking results for the discussed techniques; the comparison shown in Figure 3 can be considered as a part of the literature synthesis and not empirical research, while readers interested in quantitative benchmarking results are directed to the corresponding papers cited in Sections 2 and 5.

5. Applied Domains

The methods mentioned above have been successfully used in various fields of application, often in combination. For example, in the field of cyber security, the study of 2025 used the SHAP, LIME, and Grad-CAM approaches in combination for providing explanations to the classification of malware image by deep learning algorithm, arguing that the lack of accountability due to opaque nature of models is the problem in the area and that combination of multiple explanation methods leads to higher trust in classification of the analysts. In medical imaging, the researchers used the Kernel SHAP, LIME, and Grad-CAM approaches for providing explanation to the classification of mammographic images into categories of normality, benign and malign abnormalities using convolutional neural networks, stating that the Grad-CAM turned out to be particularly efficient for highlighting distinguishable patterns in the images depending on their categories. There is also an application in detection of pediatric obstructive sleep apnea from single-channel airflow signals, where the Grad-CAM and SHAP approaches were used, stating that the combination of both approaches provides complementary explanation with one approach (Grad-CAM) pointing at abrupt changes in the signal at apneic event onset/offset and the second (SHAP) detecting subtle noise-embedded patterns. There is also an application in natural language processing with usage of SHAP and LIME methods in combination with convolutional neural network model in educational question classification system. What can be highlighted as a common feature of all these applications is the fact that no method is deemed to be sufficiently powerful alone and multi-method triangulation (especially with gradient-based approaches for spatial/temporal localization and perturbation/game-theoretic feature attribution) is a standard approach.

Table 3. Representative Applications of XAI Methods Across Domains

Domain	Task	XAI Methods Applied	Key Finding
Cybersecurity	Malware imagery classification	SHAP, LIME, Grad-CAM	Combined methods strengthen analyst trust in opaque detection models
Medical imaging	Mammographic tumor classification	Kernel SHAP, LIME, Grad-CAM	Grad-CAM revealed distinct visual patterns across abnormality classes

Domain	Task	XAI Methods Applied	Key Finding
Clinical signal processing	Pediatric sleep apnea detection	Grad-CAM, SHAP	Methods provided complementary information on event-level vs. subtle patterns
Natural language processing	Educational question classification	SHAP, LIME (with CNNs)	Enhanced transparency of automated question-classification systems
Agriculture / ecology	Fungal species image classification	Grad-CAM, Eigen-CAM, LIME	Confirmed biologically meaningful regions of model attention

6. Open Challenges

While much progress has been made in the methodology sphere, there are certain challenges still to be addressed in the literature that has been discussed above. Firstly, the lack of a single protocol for evaluating the quality of explanations is a challenge, because although the quality of the three measures of quality has been evaluated independently, there seems to be a challenge of comparing different approaches and even the implementations of the same approach. Secondly, there are several attribution methods, such as LIME, which were demonstrated to be unstable under small perturbations of the input and parameters in sampling algorithms, so that those cannot be used in cases when the reliability of the results matters. Finally, there are methods that require calculations that are exponential in terms of the number of input features and, thus, require approximation.

Table 4. Open Challenges and Corresponding Research Directions

Challenge	Description	Emerging Response in the Literature
Lack of consensus metrics	Fidelity, stability, and comprehensibility are measured inconsistently across studies	Calls for standardized evaluation benchmarks (Doshi-Velez & Kim, 2017)
Explanation instability	Small input perturbations can produce substantially different explanations, especially for LIME	Robustness-focused variants and stability-regularized sampling schemes
Computational cost of exact SHAP	Exponential complexity in the number of features	Approximation algorithms (e.g., Kernel SHAP, tree-based SHAP)
Fidelity vs. comprehensibility tension	Technically faithful explanations may not be understandable to non-expert users	Human-subject evaluation studies of explanation comprehension and trust
Post-hoc methods as an afterthought	Explanations are computed after training rather than integrated into model design	Ante-hoc / inherently explainable architectures (e.g., DxANN)

7. Future Directions

Two emerging directions appear particularly consequential for the next phase of XAI research. First, ante-hoc explainable architectures — exemplified by recent proposals such as the Deeply Explainable Artificial Neural Network, which generates per-sample, per-feature explanations as part of the model's forward pass rather than through an external post-hoc procedure — aim to dissolve rather than manage the interpretability–performance trade-off discussed in Section 2.2. Second, the rapid adoption of large language models raises new explainability challenges distinct from those addressed by the image- and tabular-data-oriented methods reviewed in this paper; attributing a language model's output to specific input tokens, retrieved context, or internal representations at the scale of billions of parameters remains substantially less mature than the CNN- and tabular-focused methods surveyed here, and represents a priority area for future XAI research.

8. Conclusion

This paper has synthesized the rapidly expanding literature on explainable AI for deep learning into a taxonomy distinguishing intrinsic from post-hoc, local from global, and model-agnostic from model-specific explanation methods, and has developed a comparative framework clarifying the tradeoffs among leading methods including LIME, SHAP, Grad-CAM, and Integrated Gradients. The applied-domain survey demonstrates that no single method dominates in practice; instead, practitioners across cybersecurity, medical imaging, and natural language processing consistently combine complementary methods to triangulate model behavior. For researchers and practitioners selecting among XAI methods, the central implication is that method choice should be driven by the specific evaluative priorities of the application — whether model-agnosticism, computational efficiency, theoretical rigor, explanation stability, or end-user comprehensibility matters most — rather than by a presumption that any single method is universally superior. Issues of standardization in evaluation, stabilization in explanations, and the balance between fidelity and comprehension remain open areas of research interest, as does the improvement of XAI techniques to the emerging domain of large language models.

References

1. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115.
2. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE, 10(7), e0130140.
3. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
4. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), 1–42.
5. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.
6. Molnar, C. (2020). Interpretable machine learning: A guide for making black box models explainable. Lulu.com.
7. Nazim, S., Alam, M. M., Rizvi, S. S., Che Mustapha, J., Hussain, S. S., & Mohd Suud, M. (2025). Advancing malware imagery classification with explainable deep learning: A state-of-the-art approach using SHAP, LIME and Grad-CAM. PLOS ONE, 20(5), e0318542.
8. 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).
9. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (pp. 618–626).
10. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2019). Grad-CAM: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision, 128(2), 336–359.
11. Shapley, L. S. (1953). A value for n-person games. In Contributions to the Theory of Games (Vol. 2, pp. 307–317). Princeton University Press.
12. Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In Proceedings of the 34th International Conference on Machine Learning (pp. 3145–3153).
13. Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.
14. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning (pp. 3319–3328).
15. Zucker, D. (2025). Deeply explainable artificial neural network. arXiv preprint arXiv:2505.06731