Annals of Mathematics and Computer Science ISSN: 2789-7206
Vol 28 (2025) 115-126 https://doi.org/10.56947/amcs.v28.547

A SURVEY OF FEATURE ATTRIBUTION TECHNIQUES IN EXPLAINABLE AI: TAXONOMY, ANALYSIS AND COMPARISON

AMRIL NAZIR[1], EDMUND EVANGELISTA[1], SYED M SALMAN BUKHARI[2], AND RAVISHANKAR SHARMA[1]

Abstract. The feature attribution methods have become central to explainable artificial intelligence (XAI), providing critical insights into how machine learning (ML) models make individual and aggregate decisions. This survey presents a complete taxonomy of feature attribution techniques, organizing them into model-agnostic and model-specific categories while highlighting extensions such as rule-based and attention-based explanations. We analyze formal definitions of each method, mathematical formulations, application contexts, strengths, and limitations. A comparative analysis highlights key trade-offs among model flexibility, computational cost, explanation fidelity, and interpretability. In addition to theoretical perspectives, we provide practical comparisons of selected methods on benchmark tasks to guide real-world applicability. The emerging trends toward global interpretability, hybrid attribution approaches, and human-centered evaluation frameworks are discussed. This survey synthesizes current advancements and presents future directions for developing scalable, robust, and user-aligned feature attribution methods to advance responsible and transparent AI.

*Keywords.*Explainable AI, feature attribution, LIME, SHAP, interpretability, saliency maps, DeepLIFT, integrated gradients, model-agnostic methods, model-specific methods
2020 Mathematics Subject Classification. Primary 68T01; Secondary 68T07

1. Introduction

Over the last decade, explainable AI (XAI) has increasingly focused on methods that shed light on the inner workings of complex black-box machine learning (ML) models [1]. One of the most influential directions within XAI is feature attribution, which assigns an importance score to each input feature for a specific prediction [11]. The principal goal of these approaches is to clarify how much each feature has contributed to the final model output for a given test instance. Most feature attribution methods are post-hoc, applied after model training,

Date: Received: Apr 14, 2025; Accepted: Jun 2, 2025.
* Corresponding author

and local, targeting individual predictions and typically producing feature-based explanatory summaries for user inspection [2].

While local feature attribution methods focus on explaining individual model predictions, broader understanding often requires global explanations that summarize feature importance or decision behavior across an entire dataset. Overall explanations, such as global substitute models and aggregated feature importance scores, help uncover general model behavior. This can promote institutional trust and ensure adherence to ethical or regulatory frameworks.

Despite substantial progress, feature attribution remains an open research challenge. The current methods suffer from instability under input perturbations, high computational overhead for complex models, and sensitivity to feature correlations. Moreover, there is a lack of standardized evaluation criteria to assess explanation quality. A further challenge is designing explanations that are technically rigorous, intuitive, and accessible to diverse end-users. These concerns are especially critical in high-stakes domains such as healthcare, finance, and law, where transparent and accountable AI decision-making is essential.

Beyond these domains, feature attribution methods hold substantial promise across core areas of computer science. In software engineering, they support debugging and model auditing by identifying failure-inducing inputs. In cybersecurity, attribution techniques help localize anomalous patterns in intrusion detection systems. Natural language processing (NLP) applications benefit from feature attribution in tasks such as sentiment analysis, named entity recognition (NER), and question answering, where token-level relevance enhances interpretability. In computer vision, techniques such as Grad-CAM and Layer-Wise Relevance Propagation (LRP) facilitate visual validation in medical diagnostics, surveillance, and autonomous systems. These applications highlight the growing role of attribution methods in building responsible, traceable, and human-aligned AI systems.

Explanation from AI systems forms the foundation for accountability, transparency, and ultimately, trust. Without trust, the deployment of AI systems risks limited adoption and potential misuse. Recognizing this critical concern, this paper surveys the current state of practice in XAI, with a particular focus on feature attribution methods as a cornerstone of interpretable machine learning.

In this survey, we systematically organize feature attribution techniques into a comprehensive taxonomy of model-agnostic and model-specific approaches. We analyze the formal foundations, mathematical formulations, strengths, limitations, and application contexts associated with each method. Finally, we highlight key trade-offs in explainability research and propose future directions to advance the development of practical, scalable, and user-centered attribution techniques.

2. Proposed Taxonomy of Feature Attribution Methods

The feature attribution methods in XAI can be systematically organized under two primary dimensions: model-agnostic and model-specific approaches. The grouping of attribution techniques along these axes is crucial for understanding their trade-offs in generality, computational efficiency, explanation fidelity, and scalability.

The model-agnostic methods prioritize flexibility by treating the model as a black-box oracle, depending solely on input-output queries to generate explanations. In contrast, model-specific methods utilize internal gradients, reference-based comparisons, or specialized architectural structures (often within deep neural networks (DNN)) to produce more direct, architecture-aware attributions.

This organization is visually presented in Figure 1, which highlights the primary categories and subcategories of feature attribution methods. Within model-agnostic approaches, methods such as Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive exPlanations (SHAP) show local surrogate modeling and Shapley-value-based attribution, respectively. Also, high-precision rule-based methods such as Anchors [9] extend the model-agnostic paradigm by generating decision rules that reliably capture local model behavior.

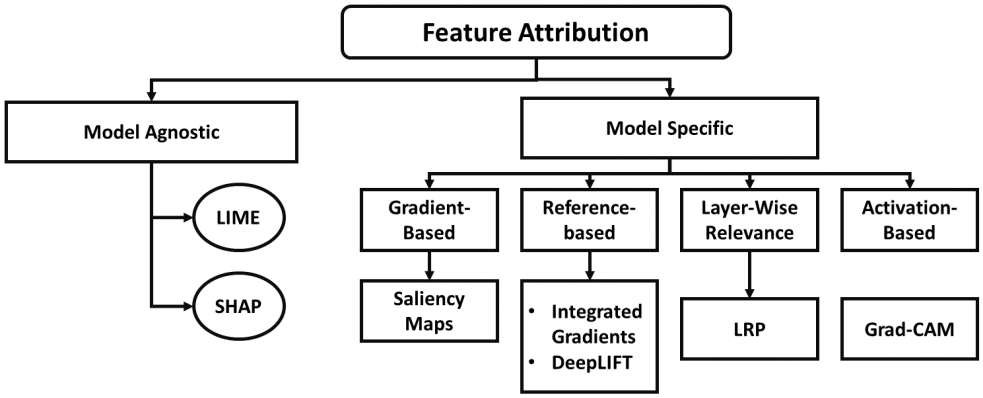

FIGURE 1. Hierarchical taxonomy of feature attribution methods, divided into model-agnostic and model-specific approaches.

The model-specific category covers gradient-based saliency maps, reference-based backpropagation methods (e.g., Integrated Gradients and DeepLIFT), Layer-Wise Relevance Propagation (LRP), Grad-CAM for convolutional neural networks (CNNs), and activation-based attention explanations used extensively in transformer models. It is important to note, however, that attention-based explanations remain a subject of debate regarding their reliability and interpretive validity [13, 14].

While Figure 1 presents a strict hierarchical organization, Figure 2 complements it by showing conceptual overlaps between model-agnostic and model-specific methods. In particular, certain gradient-based attribution techniques, such as Integrated Gradients and DeepLIFT, show characteristics that link both categories, blending theoretical properties from model-agnostic approaches with operational mechanisms from model-specific methods.

Although this taxonomy primarily focuses on local, instance-specific explanation techniques, broader global interpretability strategies, capturing general model behaviors, are also a critical area of research and are discussed in subsequent sections.

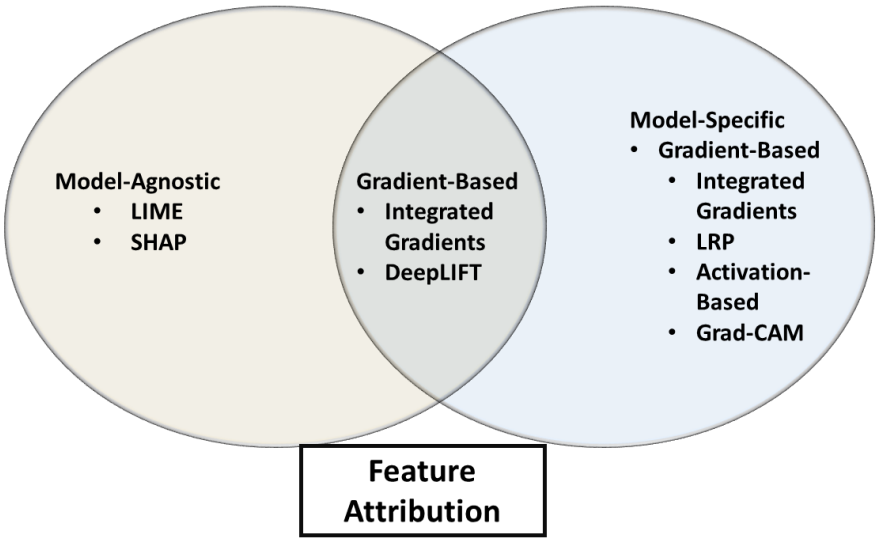

FIGURE 2. Conceptual relationships among feature attribution methods. Overlaps highlight hybrid characteristics between model-agnostic and model-specific approaches.

3. Model-Agnostic Methods

The model-agnostic feature attribution methods intend to explain model predictions without requiring access to internal model parameters or architectures. By depending solely on the input-output behavior of the model, these methods offer broad applicability across diverse model types, including black-box systems. However, this flexibility often comes at the cost of computational efficiency and may introduce stability challenges due to sampling-based approximations.

Figure 3 shows the operational workflows fundamental several major feature attribution methods discussed in the following sections, including LIME, SHAP, Saliency Maps, and Grad-CAM.

3.1. LIME: Local Interpretable Model-Agnostic Explanations.

One of the best-known model-agnostic methods is LIME [3]. LIME explains individual predictions by approximating the local decision boundary of a complex classifier or regressor with an interpretable substitute model. Formally, for a black-box model f and an instance x, LIME fits a simpler model $g \in G$ (often a sparse linear model) that optimizes a locality-weighted loss function plus a complexity penalty:

$$\xi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g),$$

where π_x defines a locality kernel around x, and $\Omega(g)$ penalizes model complexity.

In practice, LIME proceeds by randomly varying the features of x to generate a synthetic neighborhood of samples, querying the black-box model f to obtain predictions, weighting them by their similarity to x, and fitting an interpretable substitute model to approximate f locally. The feature coefficients of the substitute model are then interpreted as feature importances for the prediction.

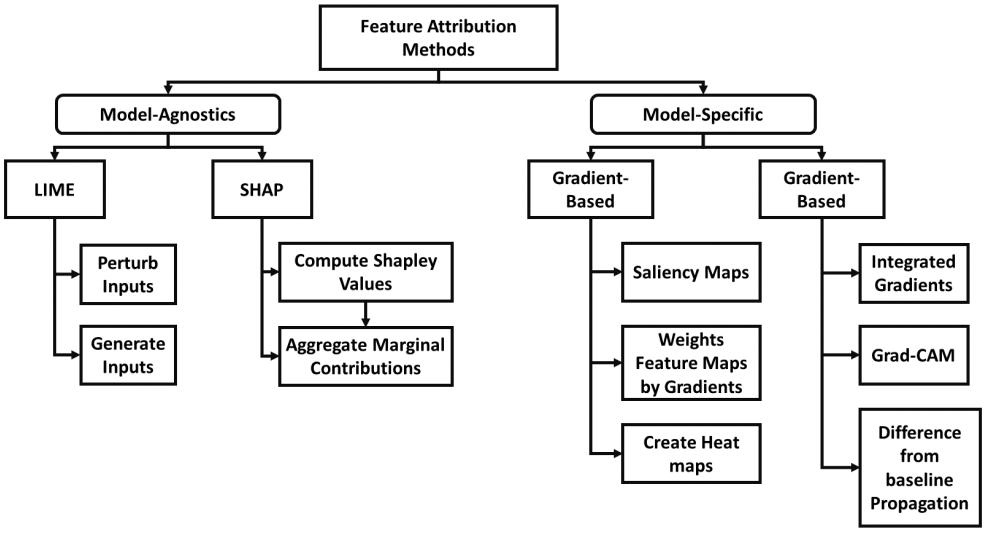

FIGURE 3. Operational Frameworks of Major Feature Attribution Methods: Processes for Model-Agnostic and Model-Specific Techniques.

LIME works well with different types of data: tables, text (by changing words), and images (by changing groups of pixels). It is considered highly interpretable, as explanations typically highlight a small set of influential features [2]. However, LIME can suffer from instability due to its random sampling strategy, and its local linear approximations may fail to capture highly nonlinear decision boundaries. Also, its computational cost grows with the complexity of the model and the number of samples needed for reliable approximation, presenting scalability challenges for large datasets and deep architectures. The extensions, such as Anchors [9], have been proposed to improve stability and precision by producing high-confidence if-then rule explanations.

Despite its popularity, evaluating the fidelity and robustness of LIME explanations remains an open challenge, motivating ongoing research into more principled evaluation metrics [15].

3.2. **SHAP: Shapley Additive Explanations.** Another widely adopted model-agnostic framework is SHAP [4], which unifies several feature attribution methods under the theoretical umbrella of Shapley values from cooperative game theory. The Shapley values provide a principled way to distribute a payout (the prediction of the model) among all input features by averaging each marginal contribution of the feature across all possible feature subsets. SHAP frames the explanation as an additive model:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i,$$

where z' is a binary vector indicating the presence or absence of features, and ϕ_i denotes the Shapley value corresponding to feature i. Formally, the Shapley

value for feature i is defined as:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(M - |S| - 1)!}{M!} \Big(f_x(S \cup \{i\}) - f_x(S) \Big),$$

where N is the complete feature set, and $f_x(S)$ denotes the model output when only features in S are present.

SHAP provides strong theoretical guarantees, including properties such as local accuracy, missingness, and consistency [2]. Multiple specialized SHAP variants have been proposed: KernelSHAP for general models via sampling, TreeSHAP for exact computations in tree ensembles [16], DeepSHAP for DNN, and LinearSHAP for linear models.

Despite its theoretical strengths, SHAP suffers from high computational complexity, particularly when applied to models with large feature spaces. The approximate techniques, such as KernelSHAP, still demand numerous model evaluations, and naive sampling can produce unrealistic feature combinations. However, optimizations such as TreeSHAP [16] have dramatically improved the feasibility of SHAP for specific model classes such as gradient-boosted trees.

Moreover, SHAP values can be sensitive to feature correlations, sometimes leading to unintuitive or misleading attributions. Addressing this sensitivity remains an active research topic. As with LIME, the objective evaluation of SHAP explanations, particularly under criteria such as stability, informativeness, and user trust, continues to present significant challenges for the explainable AI community [15].

4. Model-Specific Methods

While model-agnostic methods offer broad applicability across arbitrary model types, model-specific approaches exploit internal gradients, activations, or architectural features to produce more efficient and faithful explanations. These methods often require significantly fewer model evaluations and produce insights more directly tied to the inner workings of the model. However, their reliance on model internals limits their generalizability across different architectures. Most model-specific techniques focus primarily on local, instance-specific explanations, although emerging adaptations seek to extend them toward global interpretability.

4.1. **Gradient-Based Saliency Maps.**

Saliency maps, first introduced by Simonyan et al. [12], explain CNN-based image classifiers by computing the gradient of a specific output class score with respect to the input features. Let $S_c(x)$ be the score for class c and x_i be the i-th feature. The saliency map at feature x_i is given by

$$R_i = \frac{\partial S_c(x)}{\partial x_i}.$$

A large gradient magnitude indicates that small changes in x_i substantially affect the class score, implying that x_i is important. These gradients are visualized as heatmaps over image pixels or as importance overlays in text embeddings [12].

Saliency maps are computationally efficient, typically requiring only a single backward pass. However, they reflect only local sensitivity and are often highly noisy. Saturated regions can result in misleadingly low gradients for important features [12]. Refinements such as SmoothGrad [17] (noise-averaged gradients) and Guided Backpropagation [18] (selective gradient flow) attempt to improve clarity. Nevertheless, vanilla saliency maps lack completeness and can fail to capture global feature interactions.

4.2. **Integrated Gradients.** Integrated Gradients (IG) [5] addresses shortcomings of plain gradients by integrating the gradient along a linear path from a baseline input $x^{(0)}$ to the actual input x. For a model F and feature i:

$$\mathrm{IG}_i(x) = (x_i - x_i^{(0)}) \times \int_{\alpha=0}^{1} \frac{\partial F\big(x^{(0)} + \alpha(x - x^{(0)})\big)}{\partial x_i}\, d\alpha.$$

This integral is numerically approximated (e.g., via Riemann sums). IG satisfies completeness: the total attributions sum to $F(x) - F(x^{(0)})$. However, the robustness of the method depends on an appropriate choice of baseline, which remains an open research problem.

4.3. **DeepLIFT.** DeepLIFT [6] decomposes the difference between the output at x and at a reference $x^{(0)}$ into contributions from each input feature:

$$\sum_i C_i = F(x) - F(x^{(0)}).$$

Instead of integrating gradients, DeepLIFT propagates contribution scores using rules such as Rescale and RevealCancel, handling nonlinearities effectively. Although DeepLIFT is efficient and intuitive, it is not implementation-invariant, meaning mathematically equivalent networks can produce different attributions. Despite this, it has shown strong practical performance, particularly in sensitive domains such as genomics [6].

4.4. **Layer-Wise Relevance Propagation (LRP).** Layer-Wise Relevance Propagation (LRP) [7] traces a model's prediction backward through the network layers by distributing relevance scores. For a neuron k receiving inputs from neurons j, the relevance is computed as:

$$R_j = \sum_k \frac{z_{jk}}{\sum_{j'} z_{j'k} + \epsilon} R_k,$$

where $z_{jk} = x_j w_{jk}$ and ϵ stabilizes the denominator. LRP conserves total relevance across layers, making it interpretable and suitable for visual tasks. However, different propagation rules (e.g., α-β) and network variants can affect the resulting heatmaps.

4.5. **Grad-CAM: Gradient-Weighted Class Activation Mapping.** Grad-CAM [8] explains CNN decisions using heatmaps that highlight important image regions. It computes the gradient of the target class score for the activations A_{ij}^k

of a convolutional layer, averages these gradients globally to obtain weights α_k, and computes:

$$L^c_{\text{Grad-CAM}}(i,j) = \text{ReLU}\Big(\sum_k \alpha_k A^k_{ij}\Big).$$

Grad-CAM highlights spatial regions most influential for predictions but provides low-resolution outputs tied to the convolutional feature map's granularity. It remains architecture-specific (CNNs) and primarily emphasizes positively contributing regions.

4.6. **Attention-Based Explanations.** Attention mechanisms, prevalent in modern natural language processing (NLP) and vision models, naturally suggest interpretability by weighting input tokens or regions. It is tempting to interpret raw attention scores as feature importance indicators. However, as emphasized in Attention Is Not Explanation [13], attention weights may not reliably reflect causal influence: different attention configurations can produce identical outputs. Belinkov [10] offers a more precise view, suggesting that attention mechanisms can sometimes serve as an approximation.

Recent research proposes hybrid approaches that combine attention with attribution methods, such as gradient-attention products or attention-weighted saliency, to improve faithfulness. As exploratory tools, attention-based explanations offer insights into model behavior but should not be considered definitive proof of the model's reasoning process.

Despite their strengths, model-specific explanation methods also face challenges, including sensitivity to model modifications, instability under input changes, and the lack of standardized evaluation criteria. Developing robust, human-centered evaluation metrics remains a key area of future research.

5. Comparative Analysis and Discussion

Table 1 compares major feature attribution methods according to their model-agnosticism, explanation form, computational cost, and fidelity to the underlying model.

The trade-offs among these methods reflect fundamental tensions in XAI research: balancing interpretability, computational efficiency, and fidelity to model internals. For tabular data, additive substitute models, such as LIME and SHAP are often preferred due to their intuitive outputs. In computer vision, gradient-based and relevance propagation methods, such as Saliency Maps, LRP, and Grad-CAM, provide spatially meaningful heatmaps. In NLP, attention-based methods offer lightweight interpretability, albeit with caution regarding their reliability.

While model-agnostic methods offer broad applicability, they are often computationally expensive and sensitive to sampling variations. The model-specific methods utilize architectural insights for efficient, faithful explanations, but are tied to particular model types.

The future work should focus on integrating local and global explanations, enhancing robustness against input changes, developing standardized evaluation

TABLE 1. Comparison of feature attribution methods across key dimensions.

Method	Model-Agnostic?	Explanation Form	Comp Cost	Fidelity to Model
LIME	Yes	Local substitute model (linear)	Medium	Moderate: depends on local linearity and sampling [2]
SHAP	Yes	Shapley values	High	High: strong theoretical guarantees [2]
Saliency (Gradient)	No	Gradients as importance scores	Low	Low: sensitive to local gradients [12]
Integrated Gradients	No	Path-integrated gradients	Medium	High: satisfies completeness [5]
DeepLIFT	No	Difference-from-baseline attribution	Low	High: decomposes output differences [6]
LRP	No	Layer-wise relevance heatmap	Low	High: conserves relevance [19]
Grad-CAM	No (CNN-specific)	Spatial heatmap (positive regions)	Low	Moderate: coarse but intuitive [8]
Attention Weights	No (Transformer-specific)	Attention distribution	Very Low	Low: proxy for feature importance [13]

metrics for interpretability, and designing user-centered visualization techniques that foster trust and accessibility in AI explanations.

6. Practical Comparison of Selected Feature Attribution Methods

While theoretical properties of feature attribution methods are well-studied, practical behavior often varies depending on task characteristics, model architectures, and user expectations. To enhance the applied relevance of this survey, we briefly compare selected pairs of attribution methods based on commonly observed patterns in benchmark tasks such as text and image classification.

6.1. **SHAP vs. LIME: Model-Agnostic Local Explanations.**

SHAP [4] and LIME [3] are two of the most widely adopted model-agnostic methods for explaining individual predictions.

SHAP provides theoretically grounded attributions by computing Shapley values, ensuring properties such as local accuracy and consistency. In practice, SHAP explanations tend to be more stable across variations and better capture feature interactions, especially when applied to text or tabular classification tasks [4, 2]. However, SHAP often experiences significant computational overhead due to the combinatorial nature of Shapley value estimation, particularly in high-dimensional spaces. LIME approximates local decision boundaries using a

simpler, interpretable model, often a sparse linear regressor. While computationally efficient and providing intuitive explanations, LIME can be unstable, producing varying feature importance across repeated runs on the same instance [3, 15]. Also, the fidelity of LIME depends on the local linearity assumption, which may not hold in highly nonlinear models. On benchmark datasets such as the IMDB Sentiment Analysis [3], SHAP produces smoother, more consistent feature importance profiles than LIME, although at a greater computational cost.

6.2. **Integrated Gradients vs. DeepLIFT: Deep Model Interpretability.** IG [5] and DeepLIFT [6] are two model-specific methods designed to explain predictions of DNN through baseline comparisons.

IG computes feature attributions by integrating gradients along a straight-line path from a baseline input to the actual input, ensuring axiomatic properties such as completeness. While IG is theoretically robust, the quality of explanations depends critically on the choice of baseline; inappropriate baselines can result in misleading attributions [5].

DeepLIFT, in contrast, propagates activation differences through the network using specific rules, avoiding the need for numerical integration. It provides efficient and stable attributions but sacrifices implementation invariance: equivalent network representations may produce different explanations [6].

In image classification benchmarks such as MNIST and CIFAR-10 [17], DeepLIFT tends to generate sharper, more localized heatmaps, while IGs often produce smoother, more distributed attribution patterns. Both methods offer strong fidelity to model behavior but differ in computational demands and sensitivity to network structure.

6.3. **Practical Implications and Future Directions.** These practical comparisons show that no single feature attribution method universally dominates across tasks and domains. The researchers and practitioners must balance factors such as explanation fidelity, computational cost, stability, and interpretability requirements when selecting appropriate methods. In high-stakes applications, cross-validating insights using multiple attribution techniques is advisable to mitigate risks associated with instability or incomplete explanations. A complete empirical benchmarking of attribution methods across diverse models and datasets remains an important direction for future work, further bridging theoretical guarantees and practical usability in explainable AI.

7. Conclusion

The feature attribution methods have become essential tools for understanding the decision-making processes of machine learning (ML) models, particularly as these models are increasingly deployed in sensitive domains requiring transparency and accountability. The model-agnostic approaches, such as LIME [3] and SHAP [4], have placed the groundwork for local explanations by relying solely on input-output behavior. Meanwhile, model-specific techniques, including saliency maps, Integrated Gradients [5], DeepLIFT [6], Layer-Wise Relevance Propagation [7], and Grad-CAM [8], have utilized internal network structures

to provide more efficient and faithful attributions. The attention-based explanations [13] offer an additional interpretability perspective, though their validity remains a topic of active debate.

In practice, the choice of attribution method should be guided by domain requirements, data modality, computational constraints, and desired explanation fidelity. In high-stakes settings, cross-validating explanations using multiple methods is advisable to mitigate the risks of instability, incompleteness, or misleading attributions. As discussed in this survey, both theoretical trade-offs and practical behaviors, such as those observed in benchmark comparisons of SHAP vs. LIME and Integrated Gradients vs. DeepLIFT, must inform method selection.

While progress has been made, unifying local and global explanations remains key to complete transparency. Standardized, rigorous evaluation metrics are also needed to assess explanation fidelity, stability, and usability across diverse users and contexts. Ultimately, human-centered interpretability is paramount, ensuring explanations are accessible, actionable, and meaningful to all stakeholders.

The continued advances in attribution techniques, evaluation frameworks, and user-centered design practices will cement feature attribution as a cornerstone for building AI systems that are not only powerful but also responsible, trustworthy, and aligned with societal values.

Acknowledgement. This project was funded by a grant from Zayed University's Research and Innovation Fund to Dr. Amril Nazir. The authors are also grateful to the anonymous reviewers for their kind and helpful comments.

References

1. Hooker, S., Erhan, D., Kindermans, P. J., & Kim, B. (2019). A benchmark for interpretability methods in deep neural networks. *Advances in Neural Information Processing Systems*, **32**. DOI: 10.48550/arXiv.1806.10758
2. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI. *Information Fusion*, **58**, 82–115. DOI: 10.1016/j.inffus.2019.12.012
3. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. DOI: 10.1145/2939672.2939778
4. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, **30**. DOI: 10.48550/arXiv.1705.07874
5. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 3319–3328. DOI: 10.48550/arXiv.1703.01365
6. Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 3145–3153. DOI: 10.48550/arXiv.1704.02685
7. 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. DOI: 10.1371/journal.pone.0130140
8. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In

Proceedings of the IEEE International Conference on Computer Vision (ICCV), 618–626. DOI: 10.1109/ICCV.2017.74

9. Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, **32**(1), 1527–1535. DOI: 10.1609/aaai.v32i1.11491
10. Belinkov, Y., & Glass, J. (2024). On the explainability of attention mechanisms. *Transactions of the Association for Computational Linguistics*, **12**, 1–17. DOI: 10.1162/tacl_a_00654
11. Ali, H., Kumar, R., & Novak, M. (2023). Image-based model explanations using hybrid relevance propagation techniques. *Annals of Mathematics and Computer Science*, **19**(2), 101–120. DOI: 10.56947/amcs.v19.145
12. Singh, A., & Petrova, E. (2022). Mathematical foundations for trustworthy AI: From interpretability to verification. *Annals of Mathematics and Computer Science*, **18**(1), 67–94. DOI: 10.56947/amcs.v18.103
13. Jain, S., & Wallace, B. C. (2019). Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (pp. 3543–3556). DOI: 10.18653/v1/N19-1357
14. Wiegreffe, S., & Pinter, Y. (2019). Attention is not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 11–20). DOI: 10.18653/v1/D19-1002
15. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint* arXiv:1702.08608. DOI: 10.48550/arXiv.1702.08608
16. Lundberg, S. M., Erion, G., & Lee, S. I. (2018). Consistent individualized feature attribution for tree ensembles. *arXiv preprint* arXiv:1802.03888. DOI: 10.48550/arXiv.1802.03888
17. Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). SmoothGrad: Removing noise by adding noise. *arXiv preprint* arXiv:1706.03825. DOI: 10.48550/arXiv.1706.03825
18. Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2015). Striving for simplicity: The all-convolutional net. *arXiv preprint* arXiv:1412.6806. DOI: 10.48550/arXiv.1412.6806
19. Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K. R. (2019). Layer-wise relevance propagation: An overview. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (pp. 193–209). Springer. DOI: 10.1007/978-3-030-28954-6_10

[1] College of Technological Innovation, Zayed University, Abu Dhabi, UAE
Email address: nurmanmus@gmail.com; edmund.evangelista@zu.ac.ae; ravishankar.sharma@zu.ac.ae

[2] Capital University of Science and Technology, Islamabad, Pakistan
Email address: syedsalman.muhammad@gmail.com