A Review of Explainable AI Techniques in Machine Learning

Aishwarya Kolhe[1], Geeta Yadav[2], Dnyaneshwari Chalak[3], Dr. Sunayana Kundan Shivthare[4]
[1,2,3,4]*Department of Science & Computer Science, MAEER's MIT Arts, Commerce and Science College, Alandi (D), Pune, India*
doi.org/10.64643/ijirt.208760-459

***Abstract*—Machine learning (ML) is increasingly used in healthcare, finance, education, cybersecurity and other decision support applications. As predictive systems become more complex, it is often difficult for users to understand why a model produces a particular output. This issue is commonly described as the black box problem. Explainable Artificial Intelligence (XAI) addresses this concern by providing information that helps humans understand model predictions and behaviour. This paper reviews major XAI approaches and organizes them according to explanation scope, model dependence and the stage at which interpretability is introduced. Prominent techniques including SHAP, LIME, saliency maps, attention-based approaches, counterfactual explanations and surrogate models are discussed. Their working principles, strengths, limitations and typical application areas are compared. The paper also examines important criteria for judging explanation quality, including fidelity, stability, robustness, completeness and human interpretability.**

A key observation from the reviewed literature is that an explanation that appears convincing to a human is not necessarily a faithful representation of the underlying model. The review therefore distinguishes plausibility from faithfulness and highlights the need for systematic evaluation. Finally, research gaps involving standardized benchmarks, causal reasoning, fairness, privacy, robustness and domain-specific explanations are discussed. The review concludes that there is no single XAI method that is optimal for every model or application; the choice should depend on the model, data, decision context and needs of the intended user.

***Index Terms*—Explainable AI, Machine Learning, Model Interpretability, SHAP, LIME, Black-box Models.**

I. INTRODUCTION

Machine learning has moved from research laboratories into practical systems that support decisions in areas such as medical diagnosis, credit assessment, fraud detection, recommendation, education and cybersecurity. Modern models can identify complex patterns and often provide strong predictive performance. However, higher model complexity can make the reasoning behind an output difficult to inspect. A system may therefore provide a prediction without giving a satisfactory answer to the question: "Why did the model make this decision?"

This concern is particularly important when a prediction affects people or when a decision must be reviewed by a domain expert. For example, if a model rejects a financial application, identifies a patient as high risk, or flags an activity as suspicious, users may need to understand the factors that influenced the output.

Explainable Artificial Intelligence (XAI) has developed as a broad research area concerned with making AI behaviour more understandable to humans. Interpretability is not a single property with one universally accepted definition or measurement procedure. Doshi Velez and Kim highlighted the need for a rigorous science of interpretable machine learning and for clearer evaluation approaches [1]. XAI research consequently includes methods that expose model structure directly as well as post-hoc techniques that produce explanations after a model has been trained.

This paper reviews important XAI techniques, with emphasis on SHAP and LIME, and also considers saliency maps, attention mechanisms, counterfactual explanations and surrogate models. The objective is to organize the techniques in a practical manner, compare their characteristics and discuss the challenges that prevent any one method from being considered universally superior.

II. BACKGROUND AND RELATED WORK

The black box problem occurs when a model's mapping from inputs to outputs is difficult for a human to understand. Large neural networks and complex ensemble models can contain nonlinear interactions that are not easily expressed as simple rules. In such cases, accuracy alone is insufficient for applications where users also need transparency, error analysis or justification.

XAI methods can be divided into intrinsic and post hoc approaches. Intrinsic approaches use an interpretable model or impose constraints that make the model easier to inspect. Post hoc approaches explain an already trained model without necessarily changing its predictive structure. A second distinction is between local explanations, which focus on one prediction, and global explanations, which describe general model behaviour.

Ribeiro, Singh and Guestrin introduced LIME as a model agnostic method for explaining individual predictions through local approximation [2]. Lundberg and Lee proposed SHAP as a unified framework for additive feature attribution based on Shapley value ideas [3]. Arrieta et al. provided a broad survey of XAI concepts, taxonomies and challenges, emphasizing that explainability is connected to transparency, trust, accountability and responsible AI [4]. Molnar's work further organizes interpretable machine learning techniques and provides a practical reference for understanding feature attribution and model agnostic approaches [5].

The literature also shows that an explanation should not be judged only by whether it looks intuitive. Miller discussed explanations from a social science perspective and emphasized the importance of human expectations and context [6]. Wachter, Mittelstadt and Russell examined counterfactual explanations as a way of explaining automated decisions without requiring direct inspection of the underlying model [7].

III. TAXONOMY OF EXPLAINABLE AI TECHNIQUES

Dimension	Category	Meaning	Examples
Scope	Local	Explains one prediction	LIME, local SHAP
Scope	Global	Describes overall behaviour	Global SHAP, feature importance
Model Dependence	Model-specific	Uses model structure	Tree SHAP, saliency
Model Dependence	Model-agnostic	Works without internal model access	LIME, Kernel SHAP
Stage	Intrinsic	Interpretability built into model	Linear model, small tree
Stage	Post-hoc	Explanation generated after training	SHAP, LIME, counterfactuals

Different users ask different explanatory questions. A developer may want to understand overall feature influence, while an end user may want to know why one particular prediction was produced. Similarly, a technique designed for images may not be appropriate for a tabular dataset. Consequently, XAI selection should begin with the explanation requirement rather than with a preferred algorithm.

IV. MAJOR XAI TECHNIQUES

A. SHAP

SHAP, or SHapley Additive exPlanations, assigns contribution values to input features for a model output. Its theoretical foundation is related to Shapley values from cooperative game theory. The central idea is to estimate how the prediction changes as information from different features is considered. A positive contribution can push a prediction toward an output class, while a negative contribution can push it away.

SHAP supports local explanations for individual predictions and can also be aggregated to study global feature importance. Its main strength is the use of an additive attribution framework with useful theoretical properties. Different SHAP explainers are designed for different model families, so the choice of explainer and background data matters. SHAP values explain the behaviour of the trained model; they should not automatically be interpreted as causal effects.

B. LIME

LIME stands for Local Interpretable Model Agnostic Explanations. It explains a selected prediction by

creating perturbed samples around the input and fitting a simpler interpretable model that approximates the original model in that local neighbourhood [2]. Because it does not require detailed knowledge of the original model, LIME can be applied to many classifier types.

The principal benefit of LIME is its flexibility and intuitive local explanation. Its main limitation is sensitivity to the choice of neighbourhood, perturbation strategy and other parameters. As a result, repeated explanations can sometimes change when the input or sampling conditions change. This makes stability an important consideration when using LIME.

C. Saliency Maps

Saliency methods are commonly used with image based and deep learning models. They identify input regions that have a strong influence on an output, often using gradients or related sensitivity measures. A visual map can help an analyst see whether the model is focusing on a relevant portion of an image. However, different saliency techniques can produce different visual patterns, and a highlighted region should not automatically be regarded as a complete explanation of the model.

D. Attention-Based Methods

Attention mechanisms are common in sequence and language models. Attention values can indicate which input elements receive greater computational emphasis. They can therefore provide a useful visualization of model processing. Nevertheless, attention weights do not necessarily constitute a faithful explanation of the final prediction in every architecture. Attention should consequently be evaluated as an explanation rather than accepted as proof of model reasoning.

E. Counterfactual Explanations

Counterfactual explanations answer an actionable question: what would have to change for the model to produce a different output? For example, a financial system could indicate that a different decision might occur under an alternative combination of income, debt or other relevant attributes. Counterfactuals can be intuitive because they describe changes rather than only feature importance. Their limitation is that a mathematically valid change may not be realistic, achievable or ethically appropriate.

F. Surrogate Models

A surrogate model is a simpler model used to approximate a more complex model. A decision tree or linear model, for example, can be fitted to reproduce the behaviour of a black box system within a selected region. Surrogates can make complex behaviour easier to communicate, but their explanations are approximations. If the surrogate has low fidelity, its simplicity may come at the cost of misleading interpretation.

V. COMPARATIVE ANALYSIS

Technique	Main Strength	Main Limitation	Scope	Typical Use
SHAP	Structured feature attribution	Computationally demanding; depends on explainer/background	Local/Global	Tabular, text, image
LIME	Flexible and model-agnostic	Potential instability	Local	Tabular, text, image
Saliency	Direct visual highlighting	Method sensitivity	Local	Image/deep learning
Attention	Useful input emphasis	Not always faithful	Local/related global views	Text/sequence
Counterfactual	Action-oriented explanation	Feasibility constraints	Local	Finance, healthcare
Surrogate	Simple approximation	May lose fidelity	Local/Global	Many model types

The comparison indicates that XAI techniques answer different questions. SHAP is useful when feature contributions are required, while LIME is useful when a flexible local approximation is desired. Saliency is more natural for image inputs, whereas counterfactuals are useful when users want to

understand how an outcome could change. Surrogate models provide simplicity but must be checked for fidelity.

VI. EVALUATING EXPLANATION QUALITY

Evaluation is one of the most difficult parts of XAI because there is no single measure that captures all aspects of a useful explanation. An explanation can be technically faithful but difficult for a human to understand, or it can be very intuitive while poorly representing the actual model.

Criterion	Meaning
Fidelity	How closely the explanation represents the model's behaviour
Stability	Whether similar inputs lead to similar explanations
Robustness	Whether the explanation remains reliable under reasonable changes
Interpretability	How easily the intended user can understand it
Completeness	Whether important model behaviour is adequately represented
Computational Cost	Time and resources needed to generate explanations

Fidelity is particularly important because a persuasive explanation is not necessarily a correct explanation. Stability is also important for practical use: if a tiny change in an input causes a completely different explanation without a meaningful change in the prediction, confidence in the method can be reduced. Human interpretability is context-dependent and may require user studies rather than a purely mathematical score.

The distinction between plausibility and faithfulness is therefore central. Plausibility concerns whether an explanation appears reasonable, whereas faithfulness concerns whether it actually reflects the model's behaviour. A strong XAI evaluation should consider both rather than treating human readability as evidence of correctness.

VII. APPLICATIONS

In healthcare, XAI can help professionals inspect factors associated with predictions in diagnosis and risk assessment. In finance, explanations can support review of credit, fraud and risk decisions. In education, explanations may help teachers understand predictions about learner performance or recommendations. In cybersecurity, they can assist analysts in examining why an activity was classified as suspicious. In autonomous systems, explanations can support debugging and safety analysis.

The requirements differ across domains. A doctor may need clinically meaningful evidence, while a customer may require a concise reason for a financial decision. Therefore, explanation design should consider the intended user, consequences of error and level of technical knowledge.

VIII. CHALLENGES AND RESEARCH GAPS

Several research challenges remain. First, there is no universally accepted evaluation framework for comparing explanation quality. Second, some post hoc methods can be sensitive to parameters or small changes in the input. Third, explanations may expose associations without establishing causal relationships. A feature can be important to a model's prediction without being a cause of the real-world outcome.

Fairness and privacy also require attention. Explanations can reveal potential biases, but they can also expose sensitive information or model behaviour. Security is another concern because detailed explanations may provide information that could be exploited. Finally, a method that performs well for one data modality may not provide equally useful explanations for another.

These gaps suggest that future XAI research should focus on standardized benchmarks, robustness testing, causal reasoning, human centred evaluation and domain specific requirements. The field should move beyond simply generating explanations toward determining whether those explanations are reliable and useful in the context where they will be used.

IX. FUTURE DIRECTIONS

Future work can develop common evaluation benchmarks that combine technical and human centred measures. Causal XAI is another important direction because users often want to know not only which feature influenced a prediction but also what changes could produce a different real-world outcome. Human in the loop approaches can evaluate whether

explanations actually improve understanding and decision quality.

Domain specific XAI is also likely to become more important. Healthcare, finance, education and cybersecurity have different risks and users, so explanation requirements cannot always be identical. Research should additionally consider fairness, privacy, security and robustness as integrated properties of explainable systems.

X. CONCLUSION

Explainable Artificial Intelligence provides an important bridge between complex machine learning systems and human understanding. This review examined SHAP, LIME, saliency maps, attention-based methods, counterfactual explanations and surrogate models and classified them according to scope, model dependence and application stage.

The review shows that each technique has distinct advantages and limitations. SHAP provides structured feature attribution, LIME offers flexible local model-agnostic explanations, saliency methods support visual interpretation, attention can expose input emphasis, counterfactuals provide actionable alternatives, and surrogate models simplify complex behaviour. None of these approaches is universally best.

The most important conclusion is that XAI should be selected according to the model, data, application domain, user and explanation objective. In addition, explanation quality should be assessed using properties such as fidelity, stability, robustness and human interpretability. Future research should therefore emphasize rigorous evaluation, causal reasoning, human centred design and domain specific requirements. Such developments can help make machine learning systems more transparent, accountable and responsibly deployable.

REFERENCES

[1] F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” arXiv:1702.08608, 2017.

[2] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should I trust you?’: Explaining the predictions of any classifier,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[3] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Advances in Neural Information Processing Systems 30, 2017, pp. 4765–4774.

[4] A. B. Arrieta et al., “Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” Information Fusion, vol. 58, pp. 82–115, 2020.

[5] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2nd ed., 2022.

[6] H. Chinni, Y. K. Sharma, S. Shevkari, J. L. Vydehi, Saurabh, and M. Kavitha, “Comparative evaluation of custom CNN architectures and transfer learning models for image classification in PyTorch,” in Proc. 2026 13th Int. Conf. Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2026, pp. 1–6, doi: 10.23919/INDIACom70271.2026.11525435.

[7] S. Wachter, B. Mittelstadt, and C. Russell, “Counterfactual explanations without opening the black box: Automated decisions and the GDPR,” Harvard Journal of Law & Technology, vol. 31, no. 2, pp. 841–887, 2018.

[8] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in *Proc. 34th Int. Conf. Machine Learning*, PMLR, vol. 70, pp. 3319–3328, 2017.