International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 5, Sep-Oct-2026, ISSN (Online): 2395-566X

Explainable AI: Making Machine Learning Decisions Transparent And Interpretable — A Review Of Methods, Evaluation And Human-Centred Explanation

Padir Aditi Vinod[1]; Prof. J. M. Patade[2]; Prof. M. C. Kolse[3]; Prof. S. K. Said[4]
Department Of Artificial Intelligence And Data Science Jaihind College Of Engineering Pune, Maharashtra, India

Abstract—Machine learning models increasingly influence decisions in healthcare, finance, hiring and public services, yet many of the most accurate models are complex black boxes whose reasoning cannot be inspected. Explainable Artificial Intelligence (XAI) aims to make these decisions transparent and interpretable so that users can understand, trust, debug and contest them. This paper reviews XAI as a general discipline rather than for a single domain. It clarifies concepts of interpretability, explainability, transparency and faithfulness, and organizes methods into inherently interpretable models, global explanations such as partial dependence and accumulated local effects, local feature attribution with LIME, SHAP and integrated gradients, visual explanations for deep networks, concept-based explanations and counterfactuals. Evaluation of explanations, including sanity checks that reveal misleading saliency maps, and insights from social science on what makes a good explanation are discussed. A five-layer explainable ML pipeline comprising data, model, explanation engine, evaluation and audience-specific presentation is proposed, with research gaps, objectives, a comparison of methods and challenges such as the accuracy–interpretability trade-off, unstable explanations and regulation. The review concludes that explanations should be faithful, evaluated and designed for the people who use them.

Index Terms—Explainable AI, interpretability, SHAP, LIME, integrated gradients, saliency maps, counterfactual explanations, responsible AI

I. INTRODUCTION

A model that predicts loan default, disease risk or job suitability may be highly accurate, but if no one can explain its reasoning, errors and biases remain hidden. Caruana *et al.* [1] described a pneumonia risk model that learned asthma patients were at lower risk, because they received more intensive care in the training data; an interpretable model exposed this dangerous pattern, which a black box might have hidden.

Explainable AI addresses this problem. Doshi-Velez and Kim [2] called for a rigorous science of interpretability, and Lipton [3] warned that the term is used loosely. Surveys by Guidotti *et al.* [4], Adadi and Berrada [5] and Barredo Arrieta *et al.* [6] have organized the growing range of methods.

This paper reviews XAI concepts, methods and evaluation, and proposes a general pipeline for explainable machine learning.

II. BACKGROUND AND FUNDAMENTAL CONCEPTS

2.1 Key Concepts

Interpretability is the degree to which a human can understand a model's cause–effect behaviour; explainability refers to methods that describe it; faithfulness is how accurately an explanation reflects the model. Fig. 1 organizes XAI methods by scope (global or local) and by whether they are model-specific or model-agnostic.

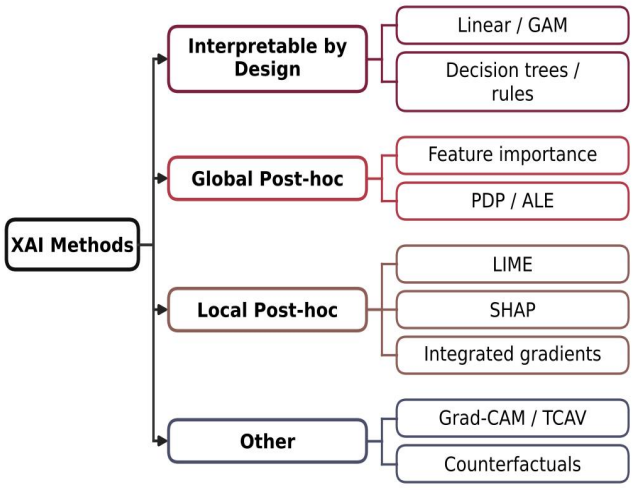


Fig. 1. Taxonomy of explainable AI methods.

2.2 Interpretable By Design

Rudin [7] argued that for high-stakes decisions, inherently interpretable models such as sparse linear models, decision lists and generalized additive models should be preferred whenever they perform comparably to black boxes.

© 2026 IJSRET

III. LITERATURE SURVEY

Global methods. Friedman [8] introduced partial dependence plots, and Apley and Zhu [9] proposed accumulated local effects, which avoid misleading results when features are correlated. Molnar [10] describes these and other methods in an accessible guide.

Local attribution. Ribeiro *et al.* [11] proposed LIME, which fits a simple model around an instance. Lundberg and Lee [12] introduced SHAP based on Shapley values, and Sundararajan *et al.* [13] proposed integrated gradients, satisfying axioms such as sensitivity and implementation invariance for deep networks.

Visual and concept-based. Grad-CAM [14] highlights image regions important for a class, and Kim *et al.* [15] proposed TCAV to test how much human-defined concepts influence predictions.

Counterfactuals and evaluation. Wachter *et al.* [16] proposed counterfactual explanations that tell users what would change a decision. Adebayo *et al.* [17] showed with sanity checks that some saliency methods produce similar maps even for randomized models, so they may not be faithful. Miller [18] drew on social science to argue that good explanations are contrastive, selective and social. Table I summarizes these works.

Ref.	Method	Key Findings	Limitations
[8]	Partial dependence plots	Global feature effects	Misleading with correlation
[1]	Intelligible GAMs in healthcare	Exposed dangerous learned pattern	Additive models only
[11]	LIME	Model-agnostic local explanations	Instability
[12]	SHAP	Consistent Shapley attributions	Computational cost
[13]	Integrated gradients	Axiomatic deep attribution	Baseline choice
[15]	TCAV	Human-concept testing	Needs concept examples
[17]	Sanity checks for saliency	Revealed unfaithful methods	Evaluation only
[18]	Social-science view of explanation	Contrastive, selective explanations	Conceptual

TABLE I Summary of Literature Survey

IV. RESEARCH GAP

G1: Different explanation methods can give conflicting explanations for the same prediction.

G2: Faithfulness is rarely measured, and some popular methods fail sanity checks.

G3: Explanations are often designed for experts rather than affected users.

G4: Explaining large language models and other foundation models remains difficult.

G5: Standards for what explanations regulations require are still emerging.

V. OBJECTIVES

- To clarify core XAI concepts and terminology.
- To review global, local, visual, concept-based and counterfactual methods.
- To examine how explanations are evaluated.
- To propose a general explainable ML pipeline.
- To discuss challenges and future directions.

VI. PROPOSED SYSTEM ARCHITECTURE

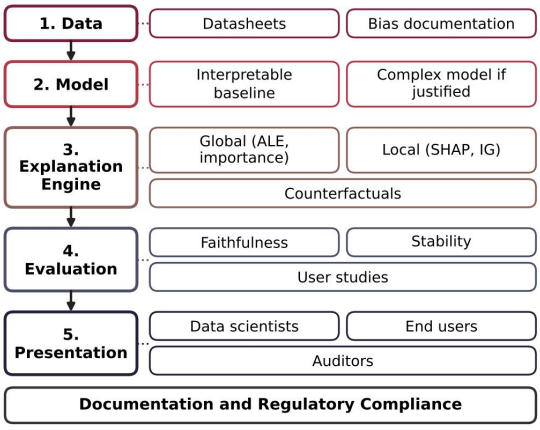


Fig. 2 presents the proposed pipeline.

Fig. 2. Proposed explainable machine learning pipeline.

© 2026 IJSRET

6.1 Data Layer

Data are documented with their sources, features and known biases, which is necessary context for any explanation.

6.2 Model Layer

An interpretable model is trained as a baseline; a complex model is used only if it is meaningfully more accurate.

6.3 Explanation Engine Layer

Global methods (feature importance, ALE) and local methods (SHAP, integrated gradients, counterfactuals) generate explanations.

6.4 Evaluation Layer

Explanations are tested for faithfulness by removing important features, for stability across similar inputs and for usefulness with users.

6.5 Presentation Layer

Explanations are tailored to the audience: technical plots for data scientists, plain-language reasons for users and documentation for auditors.

VII. METHODOLOGY AND WORKING

Fig. 3 describes the workflow. Integrated gradients attribute a prediction F(x) to feature i relative to a baseline x′:

$$IG_i(x) = (x_i - x'_i) \cdot \int_0^1 \partial F(x' + \alpha(x - x'))/\partial x_i \, d\alpha \quad (1)$$

SHAP values satisfy local accuracy, so that the prediction equals the baseline plus the sum of feature contributions, $f(x) = \varphi_0 + \Sigma \varphi_i$.

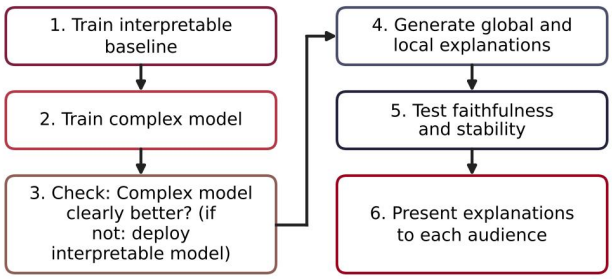


Fig. 3. Explainable ML workflow.

Faithfulness is evaluated by deletion: removing the most important features should change the prediction more than removing random ones.

VIII. COMPARATIVE ANALYSIS

Table II compares XAI methods. Interpretable models are faithful by construction, SHAP offers consistent attributions, LIME is flexible but unstable, gradient methods suit deep networks and counterfactuals are most actionable for users.

Method	Scope	Faithfulness	User Value
Interpretable model	Global + local	Exact	High
PDP / ALE	Global	High	Medium
LIME	Local	Medium	Medium
SHAP	Local + global	High	Medium–High
Integrated gradients	Local	High (axiomatic)	Medium
Counterfactuals	Local	High	Very high

TABLE II Comparison of XAI Methods

IX. CHALLENGES AND FUTURE SCOPE

Table III summarizes challenges. Future directions include explanations for LLMs, standardized evaluation benchmarks, interactive explanations and alignment of XAI with legal requirements such as those in data protection and AI regulations.

Challenge	Mitigation
Accuracy vs. interpretability	Benchmark interpretable baseline
Conflicting explanations	Use multiple methods; check agreement
Unfaithful saliency	Sanity checks; deletion tests
Non-expert users	Plain-language, contrastive reasons
Explaining LLMs	Attribution and probing research

TABLE III Challenges and Mitigation

X. CONCLUSION

Explainable AI makes machine learning decisions understandable, enabling trust, debugging and accountability. The reviewed literature offers a broad toolkit, from interpretable models and global effect plots to SHAP, integrated gradients, visual and concept explanations and counterfactuals. The key lessons are that explanations must be faithful, evaluated rigorously and designed for the people who rely on them.

10.1 Acknowledgment

The authors express their sincere gratitude to the Department of Artificial Intelligence and Data Science,

© 2026 IJSRET

Jaihind College of Engineering, Pune, for providing the facilities, guidance and encouragement needed to carry out this work. The authors also thank the faculty members and colleagues whose valuable suggestions helped improve the quality of this review.

REFERENCES

1. R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in Proc. 21st ACM SIGKDD, 2015, pp. 1721–1730.
2. F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv:1702.08608, 2017.
3. Z. C. Lipton, "The mythos of model interpretability," Commun. ACM, vol. 61, no. 10, pp. 36–43, 2018.
4. R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," ACM Comput. Surv., vol. 51, no. 5, Art. no. 93, 2018.
5. A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, 2018.
6. A. Barredo Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Inf. Fusion, vol. 58, pp. 82–115, 2020.
7. C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nat. Mach. Intell., vol. 1, no. 5, pp. 206–215, 2019.
8. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Stat., vol. 29, no. 5, pp. 1189–1232, 2001.
9. D. W. Apley and J. Zhu, "Visualizing the effects of predictor variables in black box supervised learning models," J. R. Stat. Soc. Ser. B, vol. 82, no. 4, pp. 1059–1086, 2020.
10. C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2nd ed., 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book/
11. M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?": Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD, 2016, pp. 1135–1144.
12. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. NeurIPS, vol. 30, 2017.
13. M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in Proc. ICML, 2017, pp. 3319–3328.
14. R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE ICCV, 2017, pp. 618–626.
15. B. Kim et al., "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in Proc. ICML, 2018, pp. 2668–2677.
16. S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," Harvard J. Law Technol., vol. 31, no. 2, pp. 841–887, 2018.
17. J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in Proc. NeurIPS, vol. 31, 2018.
18. T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," Artif. Intell., vol. 267, pp. 1–38, 2019.

© 2026 IJSRET