International Journal of Sustainable Development Through AI, ML and IoT
Volume 4 | Issue2 | 2025
https://ijsdai.com/index.php/IJSDAI/index
ISSN (Online): 2584-0827

Explainable Machine Learning: Methods, Applications, and Challenges

Iti Batra [1]
Vips-tc, India
Corresponding author: ***Dr. Iti Batra***

ARTICLE INFO
Received: 31 May 2025
Revised: 20 Nov 2025
Accepted: 31 Dec 2025

ABSTRACT

As machine learning models are increasingly deployed in domains such as healthcare, finance, and public policy, the opacity of high-performing models has become a central obstacle to trust, accountability, and regulatory compliance. Explainable machine learning (often discussed under the broader umbrella of explainable artificial intelligence, or XAI) addresses this obstacle by developing techniques that make model behavior interpretable to humans, either by design or after the fact. This paper surveys the current landscape of explainable machine learning. It reviews the conceptual distinction between interpretability and explainability, examines the principal families of explanation methods including intrinsically interpretable models, post-hoc model-agnostic techniques such as SHAP and LIME, and counterfactual explanations, and discusses their extension to deep learning and large language models. It further reviews applications in healthcare, finance, and other high-stakes domains, summarizes the evolving regulatory landscape shaped by the EU AI Act and related frameworks, and outlines open challenges, including the performance-interpretability trade-off, the faithfulness and stability of explanations, and the risk of a false sense of security created by explanations that appear plausible but do not reflect a model's true reasoning. The paper concludes with directions for future research, including standardized evaluation protocols and human-centered explanation design.

Keywords: *autonomous agents; agent evaluation benchmarks; AI safety and governance.*

1. Introduction

Machine learning (ML) systems now inform decisions that materially affect people's lives: which loan applications are approved, which patients are flagged for early intervention, and which job candidates are shortlisted. As these systems have grown more accurate, they have also grown more complex, with deep neural networks and large ensemble models routinely outperforming simpler statistical approaches on prediction tasks. This performance, however, has come at the cost of transparency. A gradient-boosted ensemble with thousands of trees or a neural network with millions of parameters cannot be inspected the way a linear regression coefficient table can, and stakeholders including clinicians,

loan officers, auditors, and the people subject to these decisions are left needing to trust a model's output without a clear account of how it was produced.

Explainable machine learning is the subfield concerned with closing this gap. Its goal is not necessarily to make every model interpretable in the traditional sense, but to ensure that a model's behavior can be communicated in terms that a human can understand, verify, and act upon. This paper surveys the state of the field: the technical methods used to generate explanations, the domains in which those methods are applied, the regulatory pressures driving adoption, and the unresolved challenges that limit how much trust an explanation should actually earn.

2. Background: Interpretability, Explainability, and the Black-Box Problem

The terms interpretability and explainability are often used interchangeably, but the distinction is useful. Interpretability generally refers to models whose internal structure is transparent enough that a human can trace how inputs produce outputs, sometimes called glass-box or white-box models, such as linear regression, logistic regression, decision trees, and rule lists. Explainability, by contrast, involves conveying the behavior of a model in understandable, human terms without necessarily requiring direct access to or understanding of its internal structure. Explainability can therefore be applied to interpretable models as well as to non-interpretable, black-box models, where post-hoc techniques are used to approximate or summarize the model's reasoning after training.

A deep neural network is a canonical black-box model: its millions of parameters make it effectively impossible to trace by hand, yet post-hoc techniques such as LIME or SHAP can still be used to generate human-understandable explanations of individual predictions, for example by indicating which input features most influenced a particular output. This distinction between interpretable-by-design and explainable-after-the-fact underlies much of the taxonomy used throughout the rest of this paper.

A recurring theme in the literature is the presumed trade-off between predictive performance and interpretability. Achieving state-of-the-art accuracy often requires more complex models, which tend to be harder to interpret, and finding the right balance between the two remains an open design question rather than a solved problem. Some researchers argue this trade-off is overstated for structured, tabular data, where interpretable models can match black-box performance, while others maintain that for unstructured data such as images and text, complex architectures remain necessary and explanation must happen post-hoc.

3. Methods for Explainable Machine Learning

3.1 Intrinsically Interpretable Models

The most direct route to explainability is to use models that are interpretable by construction. Linear and logistic regression models expose their reasoning through coefficients; decision trees and rule lists expose it through explicit branching logic; and generalized additive models, including the Explainable Boosting Machine (EBM), extend this transparency to non-linear relationships while retaining a feature-by-feature decomposition of the prediction. These models are attractive in regulated settings such as consumer lending, where scorecard-style models remain preferred by banking regulators precisely because their logic can be audited directly, even when black-box alternatives might offer marginally higher predictive accuracy.

3.2 Post-Hoc Model-Agnostic Methods: LIME and SHAP

When an interpretable model is not feasible, post-hoc explanation methods are applied after training to approximate the reasoning of an already-built black-box model. The two most widely adopted model-agnostic techniques are LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations).

LIME explains an individual prediction by perturbing the input around a specific instance, observing how the model's output changes, and fitting a simple, locally faithful surrogate model, typically a sparse linear model, to approximate the black box's behavior in that local neighborhood. SHAP, grounded in cooperative game theory, treats each input feature as a player in a coalition and assigns each feature a Shapley value that represents its fair contribution to the difference between the model's prediction and a baseline expectation. Because SHAP considers combinations of features rather than fitting a single local surrogate, it can provide both local, instance-level explanations and global, model-wide feature-importance summaries, and it satisfies formal consistency and local-accuracy properties that LIME does not guarantee.

Table 1 summarizes the practical differences between the two methods, which is why they are frequently used in combination: LIME for quick, intuitive, per-instance explanations, and SHAP where formal guarantees and global feature-importance rankings are required.

Table 1. Comparison of SHAP and LIME.

Criterion	SHAP	LIME
Theoretical basis	Shapley values from cooperative game theory; feature coalitions	Local linear surrogate model fit around a single instance

Criterion	SHAP	LIME
Scope	Local and global explanations	Local explanations only
Handling of non-linearity	Can capture non-linear feature interactions depending on the underlying model	Limited, since it fits a local linear approximation
Computational cost	Higher; exact computation is exponential in feature count, approximated via Kernel SHAP or model-specific variants	Lower; sampling-based perturbation and a simple regression fit
Consistency guarantees	Satisfies local accuracy, consistency, and missingness axioms	No formal consistency guarantees; explanations can vary between runs
Typical visual output	Force plots, summary/beeswarm plots, dependence plots	One feature-weight plot per explained instance

Both methods have documented limitations. SHAP's exact computation grows exponentially with the number of features, which has motivated approximations such as Kernel SHAP and model-specific variants for trees and neural networks. LIME's reliance on random perturbation and a local linear fit means that explanations can vary between runs for the same instance, and it struggles to capture non-linear feature interactions or collinear feature groups, since it interprets each coefficient as if other features were held constant, an assumption that rarely holds with correlated real-world data.

3.3 Counterfactual Explanations

Counterfactual explanations take a different approach: rather than attributing a prediction to specific features, they describe the minimal change to the input that would have produced a different outcome, for example, the change in income or credit utilization that would have flipped a loan denial to an approval. This framing, formalized in the influential work of Wachter, Mittelstadt, and Russell connecting counterfactuals to GDPR's automated-decision-making protections, is often considered actionable in a way that feature-attribution methods are not, since it tells an affected individual what they could change rather than only what mattered statistically.

Counterfactual methods have been extended in several directions: causal-discovery-based approaches that avoid assuming a fully known causal graph, sequential algorithmic recourse methods that account for the fact that a decision-making system may itself change between a rejection and reapplication, counterfactual paths through knowledge graphs, and domain-specific variants for time-series and image data. A key caveat, noted across this literature, is that a valid counterfactual assumes the underlying model remains stable; if the deployed model changes before a user acts on the recommended change, the guarantee no longer holds.

3.4 Explainability for Deep Learning and Large Language Models

Deep learning introduces additional explainability challenges because its distributed, high-dimensional representations do not map cleanly onto individual input features. Saliency and attribution methods such as DeepLIFT propagate a prediction's attribution back through network activations rather than treating the network purely as a black box queried from the outside. For transformer-based large language models, two broad strategies have emerged: attention visualization, which examines the attention weights a model assigns between tokens, and Shapley-based token attribution methods such as TokenSHAP, which estimate each token's causal contribution to an output via Monte Carlo sampling. Evaluations suggest that attention weights, while more accessible to compute, indicate what a model attended to rather than what causally drove its output, and Shapley-based token attribution has outperformed attention visualization at identifying genuinely causal tokens.

4. Applications

4.1 Healthcare

Healthcare is one of the most active application areas for explainable machine learning, since clinicians require more than a bare accuracy figure to trust a model's recommendation; they need to understand the basis for a prediction to integrate it into clinical judgment. SHAP-based analysis of diabetes prediction models, for example, has been used to identify which clinical and demographic variables most influence risk scores. Similarly, an XGBoost model for predicting heat-related illness risk from South Korean meteorological data used SHAP to identify mean daily temperature, solar radiation, and minimum temperature as the strongest contributors, supporting the model's use in early-warning public health systems where interpretability of the prediction is treated as being as important as

predictive accuracy.

4.2 Finance

In finance, explainability is closely tied to regulatory and fairness obligations. Credit scoring and loan-approval systems have historically relied on interpretable scorecard models precisely because their logic must be auditable by regulators, and counterfactual explanation methods have been proposed as a way to give applicants actionable guidance while allowing institutions to retain more complex underlying models. Explainability has also been extended to large language models used in financial analysis and reporting, where interpretability methods are being adapted to help analysts understand model-generated conclusions rather than only their tabular predecessors.

4.3 Other High-Stakes Domains

Beyond healthcare and finance, explainable machine learning has been applied to urban environmental modeling, where SHAP and LIME have been compared for predicting electromagnetic field strength from built-up area and land-use variables, and to human-computer interaction contexts such as gesture recognition systems, where explainability is treated as a design requirement rather than an afterthought in order to comply with emerging regulation. Across these domains, a common pattern emerges: explainability is most valued not simply because it improves trust in the abstract, but because it enables error detection, supports human oversight, and creates an audit trail for high-stakes automated decisions.

5. Evaluating Explanations

A persistent difficulty in this field is that there is no single agreed metric for what makes an explanation good. One influential framework distinguishes three levels of evaluation: application-grounded evaluation, where real domain experts perform real tasks using the explanations; human-grounded evaluation, using simplified tasks with lay users; and functionally-grounded evaluation, using proxy metrics that do not require human subjects at all, such as fidelity to the underlying model or the sparsity of the explanation. In practice, most published XAI methods are validated only at the functionally-grounded level, which leaves open the question of whether an explanation that scores well on a mathematical fidelity metric will actually help a clinician, loan officer, or end user make a better or fairer decision.

6. Regulatory Landscape

Regulation has become a major driver of explainability adoption. The European Union's Artificial Intelligence Act, published in the Official Journal of the EU in July 2024, is a horizontal regulatory framework that mandates transparency and explainability requirements for high-risk AI systems, including many healthcare and financial applications, alongside pre-existing obligations under the General Data Protection Regulation, whose Article 22 restricts automated decision-making with significant effects on individuals and requires safeguards such as the right to human intervention.

In practice, however, the AI Act has been criticized by domain experts for offering limited technical guidance on how explainability requirements should actually be implemented, leaving compliance obligations ambiguous and interpretability methods themselves opaque to the non-technical stakeholders who are meant to rely on them. Interview-based research with AI practitioners finds that experts view explainability as inherently context- and audience-dependent, and they call for domain-specific rules, hybrid explanation methods, and more user-centered explanation design rather than a one-size-fits-all technical mandate. Outside the EU, similar obligations are emerging elsewhere, for instance, U.S. state-level legislation such as Colorado's revised AI Act requiring deployers of automated decision-making technology to disclose how such systems inform their decisions.

In healthcare specifically, the AI Act interacts with other EU instruments including the Medical Devices Regulation and the European Health Data Space Regulation, creating a layered compliance landscape in which providers must simultaneously satisfy data protection, safety, and explainability requirements from separate but overlapping legal instruments.

7. Challenges and Open Problems

7.1 The Performance-Interpretability Trade-off

Despite ongoing debate about how absolute this trade-off really is, it remains a practical constraint in many deployments: the most accurate available model is frequently the least transparent one, forcing practitioners in high-stakes domains to choose between predictive performance and auditability, or to accept the added complexity and cost of post-hoc explanation methods layered on top of a black box.

7.2 Faithfulness and the False Sense of Security

Perhaps the most consequential open problem is that a plausible-looking explanation is not the same as a faithful one. Research on this failure mode shows that post-hoc explanation methods can be manipulated or can simply fail silently, producing explanations that look reasonable to a human reviewer while misrepresenting the model's actual decision process. This creates a false sense of security: stakeholders may over-trust a model precisely because it comes with an explanation, even when that explanation does not accurately reflect why the model produced its output. This risk is compounded by the instability noted for methods like LIME, where re-running the same explanation method on the

same instance can yield materially different feature attributions.

7.3 Quantifying Trustworthiness and Social Acceptance

Even a technically faithful explanation may fail to build trust if it does not align with a user's own intuitions or domain expertise, and there remain few standardized metrics for quantifying how interpretable a model or explanation actually is beyond proxies like model size. Social acceptance of an explanation, in other words, is a separate design problem from its technical correctness, and the two are often conflated in the literature.

7.4 Explainability at the Frontier: LLMs and Superhuman Systems

As models approach or exceed human expertise in narrow domains, explainability takes on an additional function beyond user trust: it becomes a mechanism for verifying that a system does not contain dangerous or unintended behaviors, even when evaluators have limited access to a proprietary model's internals. This verification problem, whether interpretability enables credible claims about the absence of harmful mechanisms, is described in recent interpretability research as largely unsolved, even for openly available models where full internal access exists.

8. Future Directions

Based on the trends surveyed above, several directions appear likely to shape the next phase of research in explainable machine learning:

- Standardized, task-grounded evaluation protocols that move beyond functionally-grounded proxy metrics toward application-grounded studies with real domain experts performing real decisions.
- Hybrid explanation systems that combine local and global methods, for instance pairing SHAP's global feature rankings with LIME-style or counterfactual instance-level explanations, to offset the individual weaknesses of each technique.
- Human-centered and audience-specific explanation design, tailoring explanation format and vocabulary to clinicians, regulators, or affected individuals rather than defaulting to a single technical presentation.
- Robustness research targeting the false-sense-of-security failure mode, including adversarial testing of explanation methods and stability guarantees under small input perturbations.
- Extension of causal and counterfactual methods to sequential, multi-step decision settings, where a single-shot explanation is insufficient to describe an agent's behavior over time.
- Interpretability methods purpose-built for large language models and other frontier systems, addressing both user-facing explanation and internal verification for safety-relevant behaviors.

9. Conclusion

Explainable machine learning has moved from a niche academic concern to a practical requirement across healthcare, finance, and other regulated, high-stakes domains. The field now offers a mature toolkit, ranging from intrinsically interpretable models to post-hoc methods like SHAP and LIME to counterfactual explanations, capable of producing human-readable accounts of model behavior. Yet the maturity of these tools has outpaced the maturity of their evaluation: faithfulness, stability, and genuine usefulness to the people relying on an explanation remain open and, in some cases, underexamined questions. As regulatory frameworks such as the EU AI Act push explainability from a best practice toward a legal obligation, closing this gap between explanation availability and explanation trustworthiness is likely to be one of the defining challenges for the field over the coming years.

References

Ahmed, S., Kaiser, M. S., Andersson, K., et al. (2024). A comparative analysis of LIME and SHAP interpreters with explainable ML-based diabetes predictions. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3422319

Arunika, M., et al. (2024). A survey on explainable AI using machine learning algorithms SHAP and LIME. Presented June 2024. Retrieved from ResearchGate.

Belle, V. (2025). Counterfactual explanations as plans. arXiv:2502.09205.

Dhaini, M., Ondrus, L., & Kasneci, G. (2025). From regulation to interaction: Expert views on aligning explainable AI with the EU AI Act. Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), Association for Computational Linguistics.

Frasca, M., et al. (2025). Complying with the EU AI Act: Innovations in explainable and user-centric hand gesture recognition. arXiv:2503.15528.

GLACIS. (2025). AI explainability & transparency guide 2026. Retrieved from https://www.glacis.io/guide-ai-explainability

Healthy Europe. (2026). AI in European healthcare: Promise, regulation, and reality. Retrieved from https://healthyeurope.eu/ai-european-healthcare/

Kim, J., et al. (2025/2026). Explainable machine learning for heat-related illness prediction: An XGBoost-SHAP

approach using Korean meteorological data. PMC.

Let's Data Science. (2026). SHAP vs LIME: How explainable AI actually works. Retrieved from https://letsdatascience.com/blog/shap-and-lime-making-ai-models-explainable

Pfeifer, B., Krzyzinski, M., Baniecki, H., Saranti, A., Holzinger, A., & Biecek, P. (2023). Explaining and visualizing black-box models through counterfactual paths. arXiv:2307.07764.

Salih, A. M., Raisi-Estabragh, Z., Boscolo Galazzo, I., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2024/2025). A perspective on explainable artificial intelligence methods: SHAP and LIME. Advanced Intelligent Systems, 7, 2400304. https://doi.org/10.1002/aisy.202400304

ScienceDirect / Authors. (2026). Explaining explainability: A comprehensive survey on explainable artificial intelligence and relevant industry applications. Retrieved from https://www.sciencedirect.com/science/article/pii/S2667305326000220

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. Proceedings of the International Conference on Machine Learning (ICML), PMLR, 3145-3153.

Singh, M. (2025). Interpretable AI and explainability — Part 1: Post hoc explainability. Medium.

Takahashi, D., Shimizu, S., & Tanaka, T. (2024). Counterfactual explanations of black-box machine learning models using causal discovery with applications to credit rating. arXiv:2402.02678