Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU31725

Interpretable Machine Learning Meets Statistical Inference: A Comprehensive Review of Integration Methods, Challenges, and Future Directions

Shuning Gu

School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, China
13773596633@163.com

Abstract. With the widespread deployment of machine learning models in high-stakes decision-making contexts, their inherent opacity—often termed the "black-box" problem—has raised significant concerns regarding interpretability and reliability. This paper presents a systematic and comprehensive literature review examining the convergence of interpretable machine learning and statistical inference. This paper synthesizes foundational concepts, methodological frameworks, theoretical advancements, and practical applications to elucidate how statistical tools can validate, enhance, and formalize machine learning explanations. This review critically analyzes widely adopted techniques such as SHAP and LIME, and explores their integration with statistical inference tools, including hypothesis testing, confidence intervals, Bayesian methods, and causal inference frameworks. The analysis reveals that integrated approaches significantly improve explanation credibility, regulatory compliance, and decision transparency in critical domains, including healthcare diagnostics, financial risk management, and algorithmic governance. However, persistent challenges remain in theoretical consistency, computational efficiency, evaluation standardization, and human-centered design. This paper concludes by proposing a structured research agenda focusing on unified theoretical frameworks, efficient algorithmic implementations, domain-specific evaluation standards, and interdisciplinary collaboration strategies to advance the responsible development and deployment of explainable AI systems.

Keywords: Interpretable machine learning, statistical inference, model explanation, confidence intervals, explainable AI

1. Introduction

The rapid adoption of artificial intelligence in critical fields such as healthcare, finance, and justice has fundamentally transformed decision-making processes. Machine learning models, particularly deep neural networks, demonstrate exceptional predictive performance in complex tasks. However, their growing complexity also renders decision-making opaque to human users, leading to the widely recognized "black-box" problem.

DOI: 10.54254/2755-2721/2026.GU31725

This opacity poses substantial ethical, legal, and practical challenges. In healthcare, clinicians must understand the reasoning behind AI-based diagnoses; financial regulators require explanations for automated credit decisions; and the justice system must ensure that defendants comprehend algorithmic assessments that could affect sentencing. These concerns have catalyzed the emerging field of explainable AI (XAI), which seeks to enhance the transparency, interpretability, and accountability of AI models.

At the same time, the field of statistical inference offers a mature and rigorous framework for reasoning under uncertainty. Methods such as hypothesis testing, confidence intervals, and Bayesian inference enable systematic quantification of uncertainty, hypothesis validation, and reliable conclusions from data.

The convergence of interpretable machine learning and statistical inference represents a promising frontier in AI research. Statistical inference can provide the theoretical rigor and validation mechanisms often lacking in current explainability methods, while machine learning delivers predictive capabilities beyond those of traditional statistical models. This interdisciplinary effort aims to develop AI systems that are not only accurate but also transparent, trustworthy, and ethically aligned. This paper provides a systematic review of this rapidly evolving field, covering its theoretical foundations, methodological innovations, practical applications, and emerging challenges. It focuses on four core questions: existing theoretical frameworks, methods for statistical validation, applications in high-risk domains, and future research directions. The review aims to serve as a comprehensive reference for building transparent, accountable, and ethically sound AI systems.

2. Theoretical foundations and frameworks

2.1. Interpretable machine learning: conceptual landscape

Interpretability in machine learning refers to the degree to which a human can understand, trust, and appropriately act upon a model's decisions [1]. This multidimensional concept encompasses several interrelated aspects: transparency (understanding how a model works), justifiability (providing reasons for specific decisions), informativeness (conveying relevant information), and uncertainty awareness (communicating confidence in predictions).

Approaches to interpretability are broadly categorized into two paradigms: intrinsically interpretable models and post-hoc explanation methods. Intrinsically interpretable models, such as linear regression, decision trees [2,3], rule-based systems, and generalized additive models, are designed with transparency as a primary consideration. These models typically offer direct interpretability through their structure and parameters, but may sacrifice some predictive power compared to more complex alternatives.

Post-hoc explanation methods aim to explain already-trained complex models without modifying their architecture or compromising their performance. These methods include local approximation techniques (e.g., LIME [4]), feature attribution methods (e.g., SHAP [5,6], Integrated Gradients), example-based explanations (e.g., counterfactuals [7], prototypes), and visualization tools (e.g., partial dependence plots [8], activation maximization [9]). While flexible and model-agnostic, post-hoc methods face theoretical challenges regarding faithfulness, stability, and completeness.

DOI: 10.54254/2755-2721/2026.GU31725

2.2. Statistical inference: core principles and methods

Statistical inference provides systematic approaches to conclude data while accounting for uncertainty [10]. Key components include: (1) Estimation theory: Methods for deriving point estimates (maximum likelihood, method of moments) and interval estimates (confidence intervals, credible intervals) of population parameters from sample data. (2) Hypothesis testing: Formal procedures (t-tests, ANOVA, permutation tests) for evaluating claims about population parameters based on observed evidence, with established error control mechanisms (Type I/II errors, p-values). (3) Bayesian inference: A probabilistic framework that combines prior knowledge with observed data to update beliefs, producing posterior distributions that naturally quantify uncertainty. (4) Causal inference: Methods (potential outcomes framework, structural causal models, instrumental variables) for identifying cause-and-effect relationships beyond mere correlation [4].

These statistical tools enable rigorous assessment of model stability, parameter significance, prediction uncertainty, and causal relationships—aspects often underdeveloped in standalone machine learning explanations [1,9].

2.3. Theoretical integration: motivations and frameworks

The integration of interpretable ML and statistical inference is theoretically motivated by several converging factors: (1) Epistemological alignment: Scientific knowledge traditionally requires not just predictive accuracy but also explanatory power and uncertainty quantification—standards that statistical inference formalizes. (2) Regulatory requirements: Emerging AI regulations (GDPR's "right to explanation," the Algorithmic Accountability Act, and the EU AI Act) increasingly mandate explanations that are both human-understandable and statistically validated. (3) Risk management: In high-stakes applications, understanding uncertainty and potential errors is as important as the predictions themselves. (4) Scientific validation: For AI to be integrated into scientific discovery pipelines, its outputs must withstand peer review and replication—processes grounded in statistical principles.

Theoretical frameworks for integration are emerging across multiple dimensions: game-theoretic approaches (extending SHAP with uncertainty quantification), Bayesian interpretability (treating explanations as probabilistic statements), frequentist validation (applying hypothesis testing to explanation stability), and causal interpretability (grounding explanations in causal mechanisms rather than correlations).

3. Methodological integration approaches

3.1. Statistical validation of explanations

Post-hoc explanations can be significantly enhanced through statistical validation techniques:

Uncertainty Quantification for Feature Importance: Methods like bootstrap confidence intervals for SHAP values [5] or Bayesian credible intervals for feature attribution provide measures of stability and reliability. These intervals help distinguish truly important features from those whose importance estimates fluctuate considerably with data sampling.

Statistical Testing of Explanation Significance: Hypothesis tests can determine whether explanation components (e.g., feature contributions, counterfactual changes) are statistically significant. Permutation tests, for instance, can assess whether a feature's attribution differs significantly from what would be expected under random assignment [10].

DOI: 10.54254/2755-2721/2026.GU31725

Sensitivity and Robustness Analysis: Statistical measures of explanation sensitivity to input perturbations, model variations, or hyperparameter settings provide important robustness indicators. Techniques like influence functions and maximum mean discrepancy tests quantify how explanations change under controlled variations [11].

3.2. Bayesian approaches to interpretability

Bayesian methods offer a natural paradigm for uncertainty-aware explanations through several approaches: (1) Bayesian Surrogate Models: Instead of training a single surrogate model for explanation (as in LIME), Bayesian approaches learn distributions over surrogate models, capturing uncertainty in the approximation process itself. (2) Bayesian Neural Networks with Explanation Capabilities: BNNs inherently provide uncertainty estimates for both predictions and feature attributions. Recent work has extended BNNs to produce explanations with associated credibility measures. (3) Probabilistic Graphical Models for Interpretability: Bayesian networks and other probabilistic graphical models offer transparent representations of relationships among variables, with built-in uncertainty quantification through posterior distributions.

3.3. Causal inference and explainability

Causal methods provide deeper, more actionable explanations by moving beyond correlations to identify cause-and-effect relationships: (1) Causal Feature Attribution: Methods like causal SHAP extend traditional feature attribution by accounting for causal relationships among features, addressing the feature independence assumption that often breaks down in real-world data. (2) Counterfactual Explanations with Causal Validity: Rather than generating arbitrary counterfactuals, causally valid counterfactuals respect known or discovered causal relationships, making them more meaningful and actionable. (3) Mediation Analysis for Model Understanding: Statistical mediation analysis helps decompose a model's decision process into direct and indirect effects through intermediate variables, providing nuanced explanations of how inputs affect outputs. (4) Causal Discovery for Model Interpretation: Algorithms for discovering causal structures from data can be integrated with explanation methods to ground interpretations in plausible causal mechanisms rather than mere statistical associations.

3.4. Multimodal and interactive explanation systems

Advanced explanation systems combine statistical validation with interactive interfaces and multimodal representations: (1) Uncertainty-Aware Visualization: Visualization tools that explicitly represent uncertainty in explanations—through error bars, confidence bands, or probabilistic highlighting—help users appropriately weigh explanation components. (2) Interactive Statistical Testing Interfaces: Systems that allow users to select hypotheses about model behavior and receive statistically validated tests of those hypotheses support exploratory understanding of complex models. (3) Explanation Dashboards with Statistical Diagnostics: Comprehensive interfaces that present multiple explanation perspectives alongside statistical quality metrics (stability scores, confidence measures, and consistency indicators) support more informed interpretation.

DOI: 10.54254/2755-2721/2026.GU31725

4. Domain applications and case studies

4.1. Healthcare and clinical decision support

In medical applications, the integration of statistical inference with model explanations has produced particularly valuable advances: (1) Radiology AI with statistical validation: Deep learning models for medical image interpretation now increasingly incorporate confidence intervals for both diagnoses and saliency map explanations. Bayesian deep learning approaches provide radiologists with uncertainty estimates for AI-detected anomalies, supporting more nuanced clinical decision-making. (2) Clinical risk prediction with explainable uncertainty: Models predicting patient outcomes (mortality, readmission, and complication risks) are being enhanced with statistically validated explanations that distinguish between certain and uncertain risk factors. This allows clinicians to focus intervention efforts on factors with both high importance and high confidence. (3) Drug discovery with causal explanations: In pharmaceutical research, AI models for molecule screening are being combined with causal inference methods to explain why certain compounds are predicted to be effective, moving beyond correlative features to hypothesized mechanisms of action.

4.2. Financial services and risk management

The highly regulated financial sector has pioneered several integration approaches: (1) Credit scoring with statistically-validated explanations: Financial institutions are implementing SHAP-based explanation systems with confidence intervals for feature importance. These systems not only explain credit decisions but also quantify the reliability of those explanations, supporting regulatory compliance and customer communication. (2) Algorithmic trading with uncertainty-quantified explanations: Trading algorithms increasingly incorporate explanation systems that provide real-time explanations for trading decisions with associated confidence measures. This helps traders understand when to trust automated recommendations versus exercising human judgment. (3) Fraud detection with causal attribution: Advanced fraud detection systems combine anomaly detection with causal explanation methods to not only flag suspicious transactions but also explain the causal pathways suggesting fraud, improving investigator efficiency and reducing false positives.

4.3. Public policy and algorithmic governance

Government applications present unique challenges and opportunities: (1) Social benefit allocation with transparent reasoning: Algorithms determining eligibility for social programs are being enhanced with explanation systems that provide statistically-validated reasons for decisions. This supports both fairness auditing and applicant communication. (2) Criminal justice risk assessment with calibrated explanations: Risk assessment tools in criminal justice increasingly incorporate uncertainty quantification in their explanations, helping judges and parole boards appropriately weigh algorithmic recommendations against other evidence. (3) Policy impact prediction with causal explanations: Models predicting policy outcomes are integrating causal inference methods to explain not just correlations but hypothesized causal mechanisms, supporting more robust policy design and evaluation.

4.4. Industrial and engineering applications

Technical domains present distinct interpretability requirements: (1) Predictive maintenance with reliability-explained models: AI systems predicting equipment failures are being enhanced with

DOI: 10.54254/2755-2721/2026.GU31725

explanations that quantify both the prediction confidence and the reliability of the explanation itself, supporting maintenance prioritization decisions. (2) Manufacturing process optimization with causal insights: Models optimizing complex manufacturing processes combine predictive power with causal discovery methods to explain why certain parameter adjustments improve outcomes, facilitating process understanding and control. (3) Autonomous systems with explainable decision-making: self-driving vehicles and other autonomous systems are incorporating real-time explanation systems with statistical validation, crucial for debugging, safety certification, and human oversight.

5. Current challenges

5.1. Weak theoretical foundation

Existing interpretation methods (e.g., SHAP [5]) often rely on unrealistic independence assumptions and lack reliable uncertainty quantification mechanisms [9]. They tend to over-rely on correlation-based explanations rather than causal explanations [7], limiting their role in decision support.

5.2. Low computational efficiency

Statistical validation processes (e.g., Bayesian inference, causal discovery) require significant computational resources, making it difficult to handle large-scale models and real-time scenarios. Scalability has become a critical bottleneck.

5.3. Lack of evaluation system

There is no unified standard for evaluating interpretation quality. Existing assessments overlook user cognition and decision-making effectiveness, and subjective dimensions such as comprehensibility are difficult to quantify.

5.4. Insufficient interaction design

Interpretations often contain excessive technical details, increasing cognitive load for users. Furthermore, there is a lack of personalized and context-aware presentation methods.

5.5. Ethical and governance dilemmas

Interpretations may reinforce or amplify algorithmic biases, while uncertain explanations lead to ambiguous accountability. Additionally, there is dual pressure from privacy protection and compliance requirements.

6. Future directions

Establish a unified framework that integrates statistical rigor with the flexibility of machine learning. Develop theories for modeling feature dependencies and uncertainty propagation, and strengthen cross-disciplinary research on causal explanations and deep learning theory.

Develop lightweight and scalable interpretation algorithms, explore distributed computing and approximate reasoning techniques, and build a hierarchical interpretation system for large models.

Establish a multi-level evaluation system that combines technical metrics, application standards, and user experience. Promote the development of interpretation standards for key domains.

DOI: 10.54254/2755-2721/2026.GU31725

Design adaptive and interactive interpretation interfaces. Conduct educational initiatives to improve interpretation literacy, taking into account the cognitive differences among diverse user groups.

Develop fairness-enhanced and auditable interpretation technologies. Promote the alignment of industry norms and policies, and explore agile governance models [12,13].

Establish long-term cross-disciplinary collaboration mechanisms. Cultivate multidisciplinary talent and advance the systematic development of trustworthy AI through deep cooperation.

7. Conclusion

This comprehensive review has examined the rapidly evolving intersection of interpretable machine learning and statistical inference, highlighting both significant progress and persistent challenges. The integration of these fields represents more than a technical combination—it embodies a fundamental shift toward AI systems that are not only powerful but also transparent, trustworthy, and accountable. Statistical inference provides essential tools for validating, quantifying uncertainty in, and formalizing machine learning explanations. From confidence intervals for feature importance to causal frameworks for counterfactual explanations, statistical methods enhance both the reliability and actionability of explanations. Domain applications in healthcare, finance, public policy, and industry demonstrate the practical value of these integrated approaches, particularly in high-stakes contexts where understanding uncertainty is as important as the predictions themselves.

However, challenges remain, including theoretical gaps, computational limitations, evaluation deficits, human factors issues, and ethical concerns. These interdisciplinary challenges require collaboration across computer science, statistics, domain sciences, and ethics. Looking forward, key priorities include developing unified theoretical frameworks, creating efficient algorithms for practical applications, establishing comprehensive evaluation standards, fostering interdisciplinary collaboration, and implementing technically-informed governance approaches.

As AI systems become more integrated into critical decision-making, transparent and statistically-grounded explanations will grow in importance. By advancing this research, we can work toward AI systems that are intelligent, interpretable, accurate, accountable, and aligned with human values and societal well-being.

References

[1] Lipton, Z. C. (2018). The mythos of model interpretability. Communications of the ACM, 61(10), 36–43.
[2] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. In the International Conference on Learning Representations.
[3] Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable (2nd ed.). https: //christophm.github.io/interpretable-ml-book/
[4] 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.
[5] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems, pp. 4765–4774.
[6] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.
[7] Pearl, J. (2009). Causality: Models, reasoning, and inference (2nd ed.). Cambridge University Press.
[8] 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.
[9] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the 21st ACM SIGKDD

DOI: 10.54254/2755-2721/2026.GU31725

International Conference on Knowledge Discovery and Data Mining, pp. 1721–1730.
[10] Efron, B., & Hastie, T. (2016). Computer age statistical inference: Algorithms, evidence, and data science. Cambridge University Press.
[11] Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. Journal of the Royal Statistical Society: Series B, 82(4), 1059–1086.
[12] European Union. (2016). General Data Protection Regulation (GDPR). Regulation (EU) 2016/679.
[13] Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". AI Magazine, 38(3), 50–57.