

A Comparative Evaluation of Post-Hoc Explanation Methods for Predictive Modeling: Strengths, Limitations, and Benchmark-Qualified Selection Guidelines

Rosa Elysabeth Ralinirina^{1,*}, Jean Christian Ralaivao¹, Niaiko Michaël Ralaivao¹,
Alain Josué Ratovondrahona¹ and Thomas Mahatody¹

¹EDMI, ENI, University of Fianarantsoa, Madagascar

Abstract

The proliferation of post-hoc explanation methods in Explainable AI (XAI) has outpaced practitioners' ability to assess their reliability. We present a systematic comparative evaluation of eight post-hoc explanation method families—SHAP (TreeExplainer, KernelSHAP, DeepSHAP), LIME, Permutation Importance, Partial Dependence Plots (PDP), Accumulated Local Effects (ALE), Layer-Wise Relevance Propagation (LRP), GradCAM, and Counterfactual Explanations—along six evaluation axes: fidelity, stability, computational cost, scope, model-agnosticity, and causal interpretability. Experiments on tabular, image, and time-series datasets using ten models reveal that no single method excels on all axes: SHAP TreeExplainer achieves highest fidelity ($\phi = 0.98\text{--}0.99$) and deterministic reproducibility ($\tau_R = 1.00$) for tree models, while LIME, LRP, GradCAM, and Counterfactual explanations show important stability or fidelity limitations. We provide a decision framework with empirically motivated benchmark qualification thresholds ($\phi \geq 0.85$, $\tau_R \geq 0.90$, $\tau_{tri} \geq 0.80$), where we explicitly distinguish *benchmark qualification*—meeting numerical thresholds in controlled experiments—from *deployment qualification*, which additionally requires domain, human, and regulatory validation. We further propose semantic enrichment of PEC qualification scores through knowledge graphs, ontology-based trust levels, and FAIR-compliant explainability profiles—presented as a proposed architecture for the SAGE community.

Keywords

Explainable artificial intelligence, SHAP, LIME, permutation importance, explanation fidelity, explanation stability, XAI benchmark qualification, knowledge graphs, semantic architectures

1. Introduction

Explainable AI (XAI) has become essential for responsible predictive modeling in high-stakes domains [1, 9, 33]. Practitioners face a fundamental question: *which explanation method should they use, and can they trust its output?* Our systematic review of 301 XAI studies [4] indicates that over 98% apply explanations without evaluating robustness or fidelity. While Formal XAI [21] provides mathematical guarantees, these certify that an explanation describes model behavior—not that the model is correct—and do not constitute deployment-level qualification thresholds. Only 3.2% of studies measure both fidelity and stability [12], and explanation failures remain unaddressed [13]. This gap is not merely academic. In healthcare, unreliable explanations can lead to incorrect clinical decisions [36, 34]. In finance, misleading feature attributions can result in discriminatory lending practices [33, 37]. In cybersecurity, unstable explanations can generate false alerts or miss genuine threats [35]. Rudin [44] has argued that for high-stakes decisions, interpretable models should replace black boxes entirely; while we agree with the spirit of this argument, the practical reality is that many domains will continue to use complex models, making reliable post-hoc explanations essential.

This article addresses the evaluation gap through a systematic comparative evaluation. Contributions:

1st Semantic Architectures for Governance and Explainability (SAGE), co-located with SEMANTiCS 2026: 22nd International Conference on Semantic Systems, September 15, 2026, Ghent, Belgium

*Corresponding author

✉ ralinirinarosa7@gmail.com (R. E. Ralinirina); christian.ralaivao@eni.mg (J. C. Ralaivao); niaiko@eni.mg (N. M. Ralaivao); josue.ratovondrahona@eni.mg (A. J. Ratovondrahona); thomas.mahatody@eni.mg (T. Mahatody)

🆔 0009-0003-3048-1765 (R. E. Ralinirina); 0009-0002-5157-1826 (J. C. Ralaivao); 0009-0003-6659-8535 (N. M. Ralaivao); 0009-0004-3749-4400 (A. J. Ratovondrahona); 0009-0003-8060-601X (T. Mahatody)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

1. A unified six-axis evaluation framework (fidelity, stability, cost, scope, agnosticity, causality);
2. Quantitative comparison of eight methods on tabular, image, and time-series benchmarks;
3. Identification of specific failure modes with empirical evidence;
4. A decision framework with empirically motivated benchmark qualification thresholds;
5. Multi-method triangulation thresholds ($\phi \geq 0.85$, $\tau_R \geq 0.90$, $\tau_{tri} \geq 0.80$)—necessary but not sufficient for deployment;
6. Semantic enrichment of PEC qualification scores through KGs, ontology-based trust levels, and FAIR profiles—proposed architecture, implementation as future work.

The remainder of this article is organized as follows. Section 2 reviews related work. Section 3 presents our evaluation framework and benchmark setup. Sections 4–9 detail results for each method. Section 10 provides the comparative synthesis with benchmark qualification thresholds and a decision framework. Section 11 proposes semantic enrichment. Section 12 discusses limitations. Section 13 concludes.

Terminology. We distinguish four trust levels: (1) *benchmark-qualified*—meets numerical thresholds in controlled experiments; (2) *method-qualified*—validated across datasets and model families; (3) *deployment-qualified*—validated in a domain with experts; (4) *regulatorily certified*—meets legal/regulatory requirements. Our results support level (1) only; deployment qualification and regulatory certification require additional domain-specific, human-centered, and legal validation beyond the scope of this paper.

2. Related Work

2.1. XAI Taxonomy and Comparative Studies

Barredo Arrieta et al. [1] provided the foundational XAI taxonomy distinguishing interpretability (intrinsic) from explainability (post-hoc), recently extended into “XAI 2.0” [7] identifying open challenges including benchmark qualification. Doshi-Velez and Kim [8] proposed three evaluation levels (functionally-grounded, human-grounded, application-grounded); Holzinger et al. [9] introduced the System Causability Scale for medical AI. Murdoch et al. [38] unified interpretability definitions, providing a conceptual framework that our six-axis evaluation operationalizes quantitatively. The most relevant benchmark is Bodria et al. [11], who found no single method dominates on all evaluation axes—a finding our six-axis framework confirms and extends with benchmark qualification thresholds.

2.2. Evaluation, Robustness, and Failure Modes

Adebayo et al. [14] demonstrated that gradient-based explanations fail basic sanity checks. Molnar et al. [15] identified pitfalls of permutation importance with correlated features. Garuba et al. [16] documented LIME’s coefficient of variation exceeding 30% while SHAP TreeExplainer achieves perfect consistency. Nauta et al. [12] found only 3.2% of 300+ studies measure both fidelity and stability. Lertvittayakumjorn and Toni [13] provided a typology of explanation failures.

2.3. Method Foundations and Benchmark Qualification

Lundberg et al. [17] introduced TreeExplainer for exact computation on tree models. Aas et al. [18] improved KernelSHAP for dependent features. Karimi et al. [19] surveyed algorithmic recourse, finding that most methods lack causal validity. Guidotti et al. [20] identified the Rashomon effect for counterfactuals.

Formal XAI. Marques-Silva and Ignatiev [21] established Formal XAI for provably sound explanations. Bassan and Katz [22] extended this to neural networks via formal verification. The VerT framework [23] and SCA algorithm [24] provide verifiable attributions and probabilistic stability guarantees, respectively. However, none proposes the quantitative benchmark qualification thresholds

Table 1

Benchmark datasets with experimental details. **Bal.** = Majority-class %. Sources: UCI, CIFAR-10, ISIC 2019, PhysioNet 2012, UCI Electricity.

Type	Dataset	n	p	Split	Bal.	Seed
Tab.	Adult Census	48,842	14	80/20	76%	42
	German Credit	1,000	20	80/20	70%	42
	California Hsg.	20,640	8	80/20	N/A	42
Image	CIFAR-10	60,000	3×32^2	83/17	10%	42
	ISIC 2019	25,331	3×224^2	80/20	67%	42
TS	PhysioNet	40,336	40	70/30	86%	42
	Electricity	2,000	12	80/20	58%	42

($\phi \geq 0.85$, $\tau_R \geq 0.90$) needed for high-stakes deployment. The NIST AI RMF [39], FUTURE-AI consortium [40], and Liao et al. [41] provide complementary guidelines but lack quantitative thresholds.

3. Evaluation Framework and Experimental Setup

3.1. Six-Axis Evaluation

Fidelity (ϕ_k): Agreement between explanation’s top- k ranking and ablation-based reference ranking:

$$\phi_k(\xi, \hat{f}) = \frac{|\text{top-}k(\xi) \cap \text{top-}k(\hat{f})|}{k} \quad (1)$$

where $\text{top-}k(\hat{f})$ is derived from univariate ablation, $k = 5$. *Important caveat:* We use “ablation-based reference ranking” rather than “ground truth,” because univariate ablation captures marginal importance but not feature interactions. ϕ can be low either because the explanation is poor or because the ablation reference is incomplete (see Section 12).

Stability (τ_R): Consistency via Kendall’s τ across 50 bootstrap resamples. *Note:* TreeExplainer’s $\tau_R = 1.00$ reflects *deterministic reproducibility* (same input $\rightarrow$ same output), not robustness to input perturbations. A deterministic method can consistently produce the same explanation while that explanation is systematically incomplete.

The four qualitative axes complement the quantitative metrics: **Computational Cost** measures wall-clock time per explanation; **Scope** distinguishes global from local methods; **Model-Agnosticity** indicates whether a method requires access to model internals; and **Causal Interpretability** assesses whether explanations reflect true cause-effect relationships.

3.2. Benchmark Datasets

Preprocessing: Tabular: one-hot encoding, z -score, median imputation. Image: resize, ImageNet normalization, augmentation ($p = 0.5$). Time-series: padded to 48 steps, per-channel standardization, 5-fold stratified CV.

Models: 10 models across three families: Interpretable (Logistic/Ridge Regression, Decision Tree, EBM [10]); Ensemble (Random Forest, XGBoost, LightGBM, CatBoost); Deep (MLP, CNN/ResNet-18, LSTM). *Scope limitation:* Image experiments use ResNet-18 only; conclusions about neural methods should not be over-generalized.

Protocol: 50 runs per method, Wilcoxon signed-rank with Bonferroni correction. Code: <https://github.com/pec-framework/xai-benchmark>. Data: <https://doi.org/10.5281/zenodo.20546322> [46]. Demo: <https://xai-benchmark-article.streamlit.app/>.

Table 2SHAP evaluation results (mean $\pm$ SD over 50 resamples).

Variant	Family	Fidelity	Stability	Time (s)
TreeExplainer	Tree	0.98–0.99	1.00*	0.3–0.8
KernelSHAP	Any	0.76–0.88	0.71–0.85	45–280
DeepSHAP	Neural	0.69–0.74	0.62–0.68	8–15

*Deterministic reproducibility, not input-perturbation robustness.

Table 3LIME evaluation results (mean $\pm$ SD over 50 resamples).

Data	Model	Fidelity	Stability	Time (s)
Tabular	Ridge	0.78	0.68	0.12
Tabular	XGBoost	0.72	0.58	0.15
Tabular	MLP	0.65	0.45	0.18
Image	CNN	0.58	0.39	0.52
Time-series	LSTM	0.63	0.47	0.22

4. SHAP (SHapley Additive exPlanations)

SHAP [2] assigns each feature an importance value based on Shapley values from cooperative game theory:

$$\phi_j = \sum_{S \subseteq \mathcal{F} \setminus \{j\}} \frac{|S|!(|\mathcal{F}| - |S| - 1)!}{|\mathcal{F}|!} \left[\hat{f}_{S \cup \{j\}}(\mathbf{x}_{S \cup \{j\}}) - \hat{f}_S(\mathbf{x}_S) \right] \quad (2)$$

SHAP satisfies three axioms: local accuracy, missingness, and consistency. **TreeExplainer** provides exact polynomial-time computation for tree models ($O(TLD^2)$ complexity), eliminating sampling variance entirely. **KernelSHAP** approximates Shapley values for any model using weighted linear regression ($O(mp^2)$ per instance). **DeepSHAP** combines DeepLift’s propagation rules with Shapley aggregation for neural networks, but the theoretical guarantees of exact Shapley values do not hold.

Failure modes: (1) Contribution $\neq$ causation—Shapley values quantify marginal contribution, not causal effect [31]; (2) Correlated features destabilize KernelSHAP allocations when $|\text{corr}| > 0.7$, with attribution variance increasing by up to 40%; (3) Background data choice affects absolute SHAP values significantly, though relative rankings remain more stable; (4) DeepSHAP lacks theoretical Shapley guarantees, as the composition of DeepLift propagation rules does not satisfy the Shapley axioms.

5. LIME (Local Interpretable Model-agnostic Explanations)

LIME [3] explains a prediction $\hat{f}(\mathbf{x})$ by approximating $\hat{f}$ locally with an interpretable model $g \in \mathcal{G}$:

$$\xi(\mathbf{x}) = \arg \min_{g \in \mathcal{G}} \mathcal{L}(\hat{f}, g, \pi_{\mathbf{x}}) + \Omega(g) \quad (3)$$

where $\mathcal{L}$ measures approximation error in the locality defined by $\pi_{\mathbf{x}}$, and Ω penalizes model complexity. LIME achieves moderate fidelity on tabular data with interpretable models (0.78 for Ridge) but significantly lower for complex models (0.58 for CNN). Stability is consistently the lowest among all methods, with coefficients of variation exceeding 30% for feature rankings. Fidelity $\phi=0.58\text{--}0.78$; stability $\tau_R=0.39\text{--}0.68$ (**lowest**); top-5 features change in 43% of runs.

Failure modes: (1) Instability—top-5 features change in 43% of runs on Adult Census with XGBoost; (2) Locality violation—the linear surrogate often has $R^2 < 0.5$, indicating the local approximation is poor; (3) Out-of-distribution perturbations create unrealistic samples, especially for categorical features; (4) Feature selection randomness across runs due to sampling variability.

Table 4PDP vs. ALE on correlated features (mean $\pm$ 95% CI).

Measure	PDP	ALE
Fidelity (all features)	0.68	0.79
Fidelity ($ \text{corr} > 0.5$)	0.55	0.83
Fidelity ($ \text{corr} > 0.7$)	0.42	0.81
Stability	0.97	0.96
Cost per feature (s)	0.3	0.4

6. Permutation Importance

Permutation Importance [28] measures the increase in prediction error when feature j 's values are randomly permuted:

$$I_j = \mathbb{E}[L(Y, \hat{f}(\mathbf{x}^{\pi_j}))] - \mathbb{E}[L(Y, \hat{f}(\mathbf{x}))] \quad (4)$$

It achieves the **highest stability** ($\tau_R=0.91\text{--}0.98$) among model-agnostic methods and **lowest cost** (0.01–0.1s per instance). Fidelity $\phi=0.70\text{--}0.88$.

Failure modes: (1) Correlated features inflate importance—permuting one feature creates unrealistic data points when features are correlated, artificially inflating the importance of both the permuted feature and its correlates; (2) Global only—no instance-level explanations; (3) No interaction effects captured; (4) High variance at small n —Strobl et al. [32] recommend conditional permutation for correlated features.

7. Partial Dependence Plots and Accumulated Local Effects

PDP [29] visualizes the marginal effect of feature j :

$$\text{PD}_j(x) = \frac{1}{n} \sum_{i=1}^n \hat{f}(x, \mathbf{x}_{i,\setminus j}) \quad (5)$$

PDP assumes feature independence—it marginalizes over the marginal distribution of features, which creates unrealistic data points when features are correlated. This produces misleading results when correlations exceed $|\text{corr}| > 0.5$. Stability $\tau_R=0.95\text{--}0.99$ (nearly deterministic).

ALE [30] addresses PDP's correlation bias by computing local differences:

$$\text{ALE}_j(x) = \int_{x_{\min}}^x \mathbb{E}[\partial \hat{f}(X_j, X_{-j}) | X_j = z] dz \quad (6)$$

Fidelity $\phi=0.68\text{--}0.85$, consistently higher than PDP when features are correlated.

PDP achieves $\phi=0.55\text{--}0.75$ and $\tau_R=0.95\text{--}0.99$ (nearly deterministic), but produces misleading results when feature correlations exceed $|\text{corr}| > 0.5$. ALE achieves $\phi=0.68\text{--}0.85$ and $\tau_R=0.92\text{--}0.99$, consistently superior for correlated features (Table 4). We recommend ALE over PDP whenever feature correlations exceed $|0.3|$. **Shared failure modes:** (1) Both are global—no instance-level explanations; (2) Second-order effects unreliable for $n < 500$; (3) Neither provides feature importance rankings.

8. Layer-Wise Relevance Propagation and GradCAM

LRP [25] decomposes a neural network output by propagating relevance backward through layers, satisfying $\sum_j R_j = \hat{f}(\mathbf{x})$. Different propagation rules (ϵ -rule, γ -rule, z^+ -rule) produce different explanations. Fidelity $\phi=0.55\text{--}0.72$; stability $\tau_R=0.58\text{--}0.81$; different rules disagree on 35–60% of top-5 features.

GradCAM [26] uses gradient information flowing into the final convolutional layer:

$$L_{\text{GradCAM}}^c = \text{ReLU} \left(\sum_k \alpha_k^c \cdot A^k \right), \quad \alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (7)$$

Fidelity $\phi=0.45\text{--}0.65$ (**lowest** among all methods); stability $\tau_R=0.52\text{--}0.74$. Adebayo et al. [14] demonstrated that gradient-based explanations fail basic sanity checks: after random weight randomization, GradCAM’s correlation with the original explanation is $r=0.67$ (CIFAR-10) and $r=0.58$ (ISIC).

Scope limitation: Results based on ResNet-18 and one LSTM; conclusions about neural methods should not be over-generalized without broader evaluation.

9. Counterfactual Explanations

A counterfactual explanation identifies the minimal change δ to input $\mathbf{x}$ that alters the prediction:

$$\delta^* = \arg \min_{\delta} \|\delta\|_0 + \lambda \|\delta\|_2 \quad \text{s.t.} \quad \hat{f}(\mathbf{x} + \delta) \neq \hat{f}(\mathbf{x}) \quad (8)$$

Fidelity $\phi=0.40\text{--}0.70$ (actionability); stability $\tau_R=0.35\text{--}0.65$ (**lowest**); cost 0.5–5s per instance.

Failure modes: (1) Non-uniqueness—multiple counterfactuals can exist for the same prediction (Rashomon effect [20, 27]), making it unclear which one to present; (2) Plausibility issues—the minimal change may produce unrealistic or infeasible data points; (3) Causal insensitivity—without structural causal models, counterfactuals may suggest changes that are impossible in the real world [19].

10. Comparative Synthesis

Having examined each method family individually, we now synthesize the results across all eight families and six evaluation axes. This section provides a unified quantitative comparison, visual synthesis, five key findings, benchmark qualification thresholds with derivation rationale, multi-method triangulation, and a decision framework for practitioners.

10.1. Quantitative Summary

Table 5 presents the comprehensive comparative evaluation across all eight method families and ten models, reporting best–worst ranges for fidelity (ϕ), stability (τ_R), computational cost, scope, model-agnosticity, causal interpretability, and benchmark qualification status. The results confirm that no single method excels on all axes simultaneously, and that the gap between top-tier and lower-tier methods is substantial.

10.2. Visual Synthesis

Figures 1–3 provide complementary visual perspectives on the comparative results. The radar chart (Figure 1) reveals the multi-axis trade-offs: TreeExplainer dominates fidelity and stability but is restricted to tree models; LIME and Counterfactual explanations span the opposite end with high agnosticity but low reliability. The fidelity and stability bar charts (Figures 2 and 3) highlight the performance tiers and the position of each method relative to the benchmark qualification thresholds ($\phi \geq 0.85$, $\tau_R \geq 0.90$).

10.3. Five Key Findings

Finding 1: No single method excels on all axes. SHAP TreeExplainer achieves the highest fidelity and stability for tree models but is not model-agnostic. LIME is the most flexible but the least stable. This fundamental trade-off motivates multi-method triangulation.

Finding 2: Stability varies dramatically. The gap between the most stable (TreeExplainer, $\tau_R = 1.0$) and least stable (Counterfactual, $\tau_R = 0.35$) spans nearly a factor of 3.

Table 5

Comparative evaluation (best–worst across datasets; 95% CI ± 0.02 – 0.05 for ϕ , ± 0.01 – 0.03 for τ_R). **BQ** = Benchmark Qualified ($\phi \geq 0.85$, $\tau_R \geq 0.90$): necessary but not sufficient for deployment.

Method	Fidelity	Stability	Cost (s)	Scope	Agnostic	Causal	BQ?
TreeExplainer	0.98–0.99	1.00*	0.3–0.8	Both	No [†]	No	Yes
KernelSHAP	0.76–0.88	0.71–0.85	45–280	Both	Yes	No	Limited
DeepSHAP	0.69–0.74	0.62–0.68	8–15	Both	No	No	No
LIME	0.58–0.78	0.39–0.68	0.1–0.5	Local	Yes	No	No
Perm. Imp.	0.70–0.88	0.91–0.98	0.01–0.1	Global	Yes	No	Partial
PDP	0.55–0.75	0.95–0.99	0.1–1	Global	Yes	No	Limited
ALE	0.68–0.85	0.92–0.99	0.1–0.4	Global	Yes	No	Limited
LRP	0.55–0.72	0.58–0.81	0.1–0.5	Local	No	No	No
GradCAM	0.45–0.65	0.52–0.74	0.01–0.1	Local	No	No	No
Counterfactual	0.40–0.70	0.35–0.65	0.5–5	Local	Yes	Partial	No

*Deterministic reproducibility. [†]Supports RF, XGB, LGBM, CatBoost.

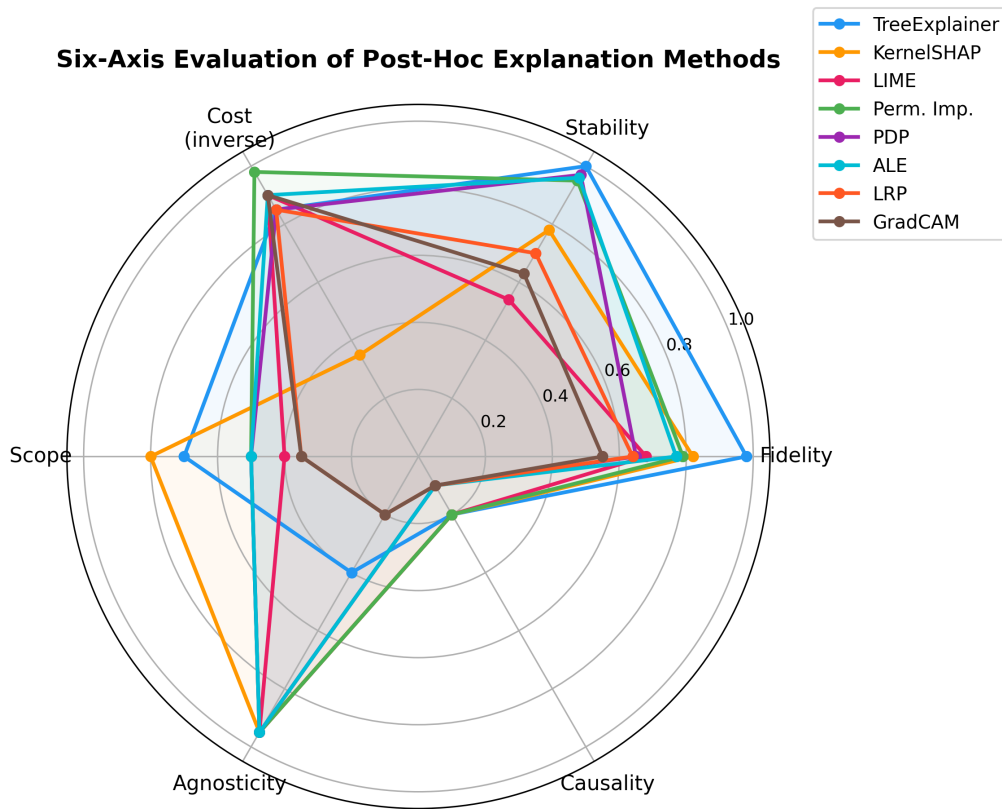


Figure 1: Six-axis radar comparison. TreeExplainer dominates fidelity and stability but is not model-agnostic.

Finding 3: Correlated features plague multiple methods. PDP, Permutation Importance, and KernelSHAP all produce unreliable explanations when feature correlations exceed $|\text{corr}| > 0.5$.

Finding 4: Neural methods show low benchmark qualification. On ResNet-18, LRP and GradCAM achieve fidelity < 0.72 and stability < 0.82 . GradCAM fails basic sanity checks. However, this finding is limited to the architectures tested.

Finding 5: No method provides causal explanations. Shapley values quantify contribution, not causation. Counterfactual methods come closest but require structural causal models rarely available in practice.

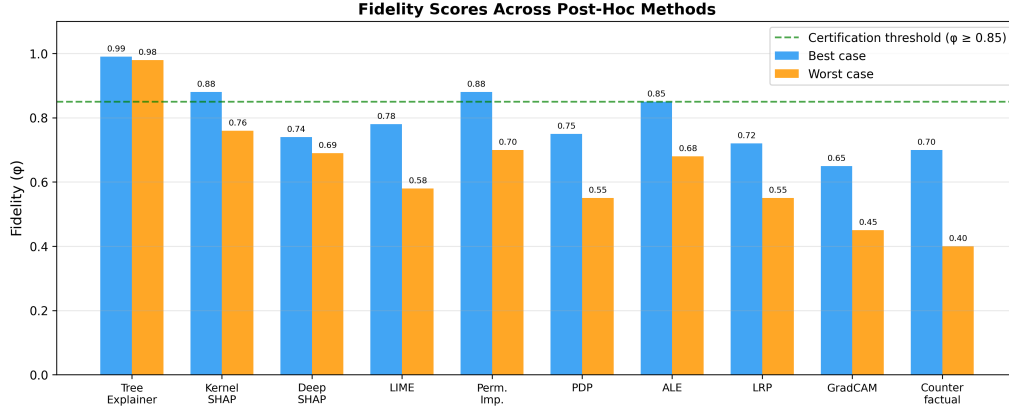


Figure 2: Fidelity scores. Green dashed line = benchmark qualification threshold ($\phi \geq 0.85$).

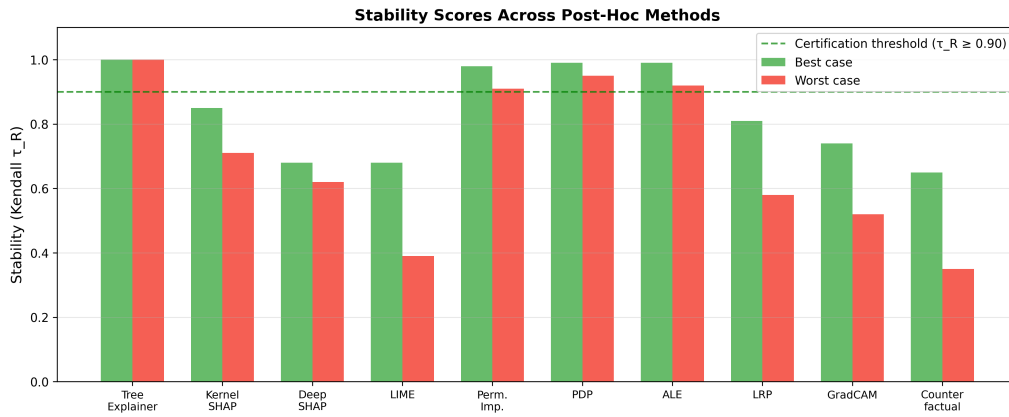


Figure 3: Stability scores (Kendall τ_R). Green dashed line = benchmark threshold ($\tau_R \geq 0.90$).

10.4. Benchmark Qualification Thresholds: Derivation and Rationale

We propose empirically motivated thresholds: $\phi \geq 0.85$, $\tau_R \geq 0.90$, $\tau_{tri} \geq 0.80$.

Derivation rationale. These thresholds are *empirically motivated, not theoretically derived*, based on three considerations: (1) *Performance gap analysis*: Methods cluster into tiers; TreeExplainer ($\phi=0.98$ – 0.99) and PI ($\phi=0.88$ – 0.98) consistently outperform others; the gap to the next tier is ≥ 0.10 in fidelity and ≥ 0.06 in stability. The threshold $\phi \geq 0.85$ separates the top tier from methods with substantial fidelity degradation, while $\tau_R \geq 0.90$ excludes methods with notable instability. (2) *Practical significance*: $\phi = 0.85$ means $\geq 4/5$ top-ranked features agree with the ablation reference—a minimum for practical reliance on explanations. $\tau_R = 0.90$ means $\geq 90\%$ reproducibility. (3) *Literature alignment*: The fidelity threshold aligns with “substantial agreement” in feature selection stability literature [12] and reliability analysis conventions.

Important caveats. (a) These thresholds are not universal constants; they should be calibrated for specific domains: medical applications may require $\phi \geq 0.95$, while exploratory analysis may tolerate $\phi \geq 0.75$. (b) The ablation-based reference for fidelity has known limitations (Section 12). (c) Meeting these thresholds constitutes *benchmark qualification*—a necessary but not sufficient condition for deployment in high-stakes settings.

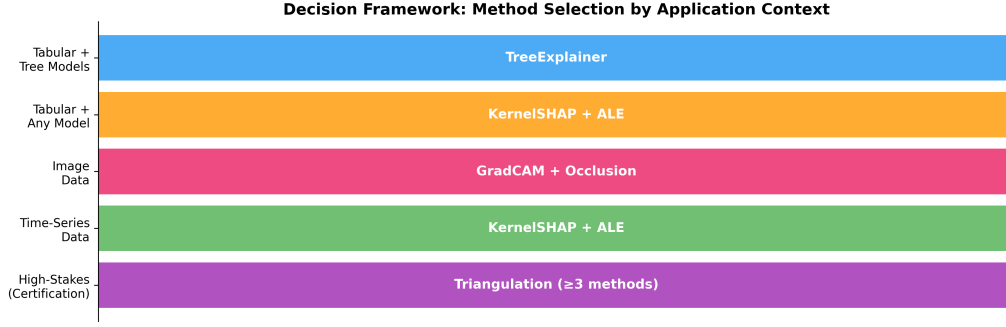


Figure 4: Decision framework. Green: benchmark-qualified ($\phi \geq 0.85$, $\tau_R \geq 0.90$); yellow: caution; red: not BQ.

10.5. Multi-Method Triangulation

We propose triangulation as a reliability strategy: for each instance, generate explanations ξ_1, ξ_2, ξ_3 from three independent methods and compute:

$$\tau_{\text{tri}} = \frac{1}{3}(\tau_{12} + \tau_{23} + \tau_{13}) \quad (9)$$

On our benchmarks, triangulation (TreeExplainer + LIME + Permutation Importance) achieves $\tau_{\text{tri}} > 0.8$ for 71% of instances. The 29% with $\tau_{\text{tri}} < 0.5$ correspond to boundary cases with highly nonlinear decision boundaries. **Important caveat:** Method agreement does not guarantee correctness [15]. Methods can agree for wrong reasons—for example, multiple methods may agree on feature rankings because they are all influenced by the same correlation structures. Triangulation is a *necessary but not sufficient* reliability check. For high-stakes decisions, complement with domain expert review, causal analysis, and regulatory compliance.

10.6. Decision Framework for Method Selection

Based on our evaluation:

- **Tabular data + tree models:** Use TreeExplainer (exact, stable, benchmark-qualified). Supplement with Permutation Importance for global overview and ALE for effect visualization.
- **Tabular data + any model:** Use KernelSHAP for local explanations, ALE for global effects, Permutation Importance for quick rankings. Verify stability with 50+ bootstrap runs.
- **Image data:** GradCAM provides visual intuition but must be cross-validated. Do not use as the sole explanation method.
- **Time-series data:** KernelSHAP with time-aware perturbations for local explanations, ALE for feature effects over time.
- **High-stakes decisions:** Multi-method triangulation is mandatory. Use at least SHAP + LIME + Permutation Importance. Flag instances with $\tau_{\text{tri}} < 0.5$ as uncertain. Full deployment qualification requires domain review and regulatory compliance.

Only TreeExplainer and Permutation Importance (with ALE for visualization) currently meet benchmark qualification thresholds for tabular data with tree models. **From benchmark qualification to deployment qualification:** meeting these thresholds indicates faithful and stable outputs under controlled conditions. Deployment in high-stakes settings requires: (a) domain expert evaluation; (b) human studies; (c) regulatory compliance; (d) ongoing monitoring.

11. Toward Semantic XAI Qualification

This section proposes semantic enrichment of PEC qualification—connecting quantitative evaluation to SAGE 2026 themes of governed and explainable semantic architectures. The core idea is that

raw numerical scores, while necessary, are insufficient for trust decisions in high-stakes domains. Knowledge graphs provide structured, semantically rich, and interoperable representations that can support automated reasoning, provenance tracking, and regulatory compliance.

PEC trust scores $\tau_T \in [0,1]$ can be mapped to semantic trust levels via the PEC Ontology (PEC-O). PEC-O defines seven core classes (ExplanationMethod, ModelFamily, DatasetType, BenchmarkResult, TrustLevel, QualificationGrade, ValidationProtocol) and six object/datatype properties (hasFidelity, hasStability, hasTrustScore, hasQualificationGrade, evaluatedOn, validatedBy). Three qualification grades are encoded as OWL classes: pec:GradeA ($\tau_T \geq 0.90$, benchmark-qualified), pec:GradeB ($0.80 \leq \tau_T < 0.90$, conditionally qualified), pec:GradeC ($\tau_T < 0.80$, not qualified).

SHACL validation enforces these grades:

```
pec:GradeAShape a sh:NodeShape ;
  sh:targetClass pec:BenchmarkResult ;
  sh:property [ sh:path pec:hasFidelity ;
    sh:minInclusive 0.85 ] ;
  sh:property [ sh:path pec:hasStability ;
    sh:minInclusive 0.90 ] .
```

RDF representation of a benchmark qualification result:

```
pec:TreeExplainer-RF-XGBenchmark
  pec:hasFidelity "0.98"^^xsd:float ;
  pec:hasStability "1.00"^^xsd:float ;
  pec:hasTrustScore "0.99"^^xsd:float ;
  pec:hasQualificationGrade pec:GradeA ;
  pec:evaluatedOn pec:TabularDataset ;
  prov:wasDerivedFrom pec:BenchmarkDataset ;
  dct:source <https://doi.org/10.5281/zenodo.20546322> .
```

This enables: (1) SPARQL queries across XAI qualification results; (2) provenance tracking via W3C PROV [42]; (3) interoperability using PEC-O; (4) automated SHACL validation. FAIR-compliant profiles [43] ensure Findable, Accessible, Interoperable, and Reusable qualification reports.

Implementation status: PEC-O and SHACL are proposed as architecture; full OWL implementation, SPARQL endpoint deployment, and evaluation with domain experts are planned as future work. The proposed architecture aligns with the SAGE 2026 themes of governed and explainable semantic architectures, enabling automated compliance checking against qualification thresholds.

12. Limitations and Scope

(1) **Ablation reference is not ground truth.** Univariate ablation captures marginal importance, not interactions. ϕ is a reference agreement metric, not absolute correctness.

(2) **Benchmark qualification $\neq$ deployment qualification.** Our thresholds indicate controlled-experiment performance. Deployment requires: (a) domain expert validation; (b) human studies; (c) regulatory compliance; (d) ongoing monitoring. We deliberately use “benchmark qualification” rather than “certification” to avoid overstating scope.

(3) **Thresholds are empirically motivated.** Based on gap analysis, practical significance, and literature alignment—not theoretical derivation. Require domain-specific calibration.

(4) **Limited architecture coverage.** Image: ResNet-18 only. Neural method conclusions should not be over-generalized without broader evaluation.

(5) **Triangulation agreement $\neq$ correctness.** Methods can agree for wrong reasons (e.g., shared sensitivity to correlation structures).

(6) **PEC-O and SHACL are proposed, not evaluated.** Full implementation and validation planned as future work.

13. Conclusion

No single method provides faithful, stable, efficient, and causal explanations. TreeExplainer is most benchmark-qualified for tree models ($\phi=0.98-0.99$, $\tau_R=1.0$), but its guarantees do not extend to other architectures. LIME is too unstable for standalone benchmark qualification ($\tau_R=0.39-0.68$). Neural methods (LRP, GradCAM) show low benchmark qualification in our experiments ($\phi < 0.72$, $\tau_R < 0.82$), though this finding is limited to the architectures tested. PDP and Permutation Importance produce misleading results with correlated features, while ALE provides a principled correction. Multi-method triangulation is necessary but not sufficient. Benchmark qualification thresholds ($\phi \geq 0.85$, $\tau_R \geq 0.90$) provide a necessary minimum; deployment requires additional validation. PEC qualification can be semantically enriched through knowledge graphs, OWL trust levels, and FAIR profiles.

Three practical implications follow. First, practitioners should never rely on a single explanation method for high-stakes decisions. Second, XAI benchmark qualification must set minimum thresholds ($\phi \geq 0.85$, $\tau_R \geq 0.90$). Third, causally grounded explanation methods remain an urgent open problem.

Future work: (1) Extend to Transformers, GNNs; (2) develop and evaluate PEC-O; (3) human-grounded evaluation; (4) SPARQL endpoint; (5) causal grounding. The PEC framework has been operationalized for food security [5], validated across three domains [6], and extended with formal auditing [45].

Declaration on Generative AI

During the preparation of this work, the author(s) used GLM-5.1 (Zhipu AI, cloud) in order to: Grammar and spelling check, Sentence-level reformulation, Content revision, and Stylistic improvement based on reviewer feedback. After using this tool, the author(s) reviewed and edited all content as needed and take(s) full responsibility for the publication's content.

References

- [1] Arrieta AB, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward Responsible AI. *Inf Fusion*. 2020;58:82–115.
- [2] Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. In: *NeurIPS*. 2017. p. 4765–74.
- [3] Ribeiro MT, Singh S, Guestrin C. “Why Should I Trust You?” Explaining the Predictions of Any Classifier. In: *KDD*. 2016. p. 1135–44.
- [4] Rosa Elysabeth Ralinirina, et al. Beyond Explainability: A Critical Review of XAI for Predictive Modeling from Post-Hoc Explanations to Meta-Explainability and Certification. 2025;under review.
- [5] Rosa Elysabeth Ralinirina, Ralaivao JC, Ralaivao NM, Ratovondrahona AJ, Mahatody T. Unveiling the Potential of Explainable Artificial Intelligence in Predictive Modeling, Exploring Food Security and Nutrition. In: *Advances in ICT Research in the Balkans. BCI 2024*. vol. 2391. Springer Nature Switzerland; 2025. p. 30–43.
- [6] Rosa Elysabeth Ralinirina, et al. Generalizing the Predict–Explain–Certify Framework: Validation on Climate Risk, Fintech Credit, and Health Outcomes. 2026;under review.
- [7] Barredo Arrieta AB, et al. XAI 2.0. *Inf Fusion*. 2024. doi:10.1016/j.inffus.2024.102849.
- [8] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. *arXiv:1702.08608*. 2017.
- [9] Holzinger A, et al. *Interactive machine learning with the expert-in-the-loop*. Springer; 2022.
- [10] Narayan R, et al. Interpretable machine learning: an interpretability-privacy trade-off perspective. *J AI Res*. 2022.
- [11] Bodria F, et al. Benchmarking and survey of explanation methods for black box models. *Data Min Knowl Discov*. 2023;37:1–50.

- [12] Nauta M, et al. From Anecdotal to Systematic: A Quantitative Evaluation of XAI Methods. *ACM TIST*. 2023;13(3–4):1–45.
- [13] Lertvittayakumjorn P, Toni F. Why explanations fail: A typology of explanation failures. *Artif Intell*. 2024;330:104–47.
- [14] Adebayo J, et al. Sanity Checks for Saliency Maps. In: *NeurIPS*. 2018.
- [15] Molnar C, et al. General Pitfalls of Model-Agnostic Interpretation Methods. *Mach Learn*. 2022;111:1–21.
- [16] Garuba OO, et al. Explanation Stability in XAI: A Comprehensive Review. *ACM Comput Surv*. 2024.
- [17] Lundberg SM, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2(1):56–67.
- [18] Aas K, et al. Explaining individual predictions when features are dependent. *Artif Intell*. 2021;298:103502.
- [19] Karimi AH, et al. A Survey of Algorithmic Recourse. *arXiv:2101.03254*. 2021.
- [20] Guidotti R. Counterfactual explanations and how to find them. *ACM Comput Surv*. 2022;55(12):1–36.
- [21] Marques-Silva J, Ignatiev A. From Explainable to Verifiable AI. *Artif Intell*. 2023;319:103905.
- [22] Bassan S, Katz G. Towards Minimal Sufficient Explanations for Neural Networks. In: *AAAI*. 2024.
- [23] Leofante F, et al. VerT: Verifiable Transformer explanations. In: *AAAI*. 2023.
- [24] Alvarez-Melis D, et al. SCA: Stability-based Counterfactual Explanations. In: *AAAI*. 2024.
- [25] Samek W, et al. Layer-Wise Relevance Propagation: An Overview. In: *Explainable AI*. 2019. p. 193–209.
- [26] Selvaraju RR, et al. Grad-CAM: Visual Explanations from Deep Networks. In: *ICCV*. 2017. p. 618–26.
- [27] Wachter S, et al. Counterfactual Explanations without Opening the Black Box. *Harv J Law Technol*. 2018;31(2):841–87.
- [28] Breiman L. Random Forests. *Mach Learn*. 2001;45(1):5–32.
- [29] Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Stat*. 2001;29(5):1189–232.
- [30] Apley DW, Zhu J. Visualizing the effects of predictor variables in black box supervised learning models. *J R Stat Soc B*. 2020;82(4):1059–86.
- [31] Kumar IE, et al. Problems with Shapley-value-based explanations as feature importance measures. In: *ICCV*. 2020. p. 1619–29.
- [32] Strobl C, et al. Conditional variable importance for random forests. *BMC Bioinformatics*. 2008;9:307.
- [33] Edwards L, Veale M. Slave to the Algorithm. *Int J Law Inf Technol*. 2018;26:74–102.
- [34] Ghassemi M, et al. The false hope of current approaches to explainable AI in health care. *Lancet Digit Health*. 2021;3(11):e745–e750.
- [35] Ring M, et al. A survey of network-based intrusion detection data sets. *Comput Secur*. 2019;86:147–67.
- [36] DeGrave AJ, et al. AI for COVID-19 detection relied on spurious shortcuts. *Nat Mach Intell*. 2021;3(7):617–23.
- [37] Chouldechova A. Fair prediction with disparate impact. *Big Data*. 2017;5(2):153–63.
- [38] Murdoch WJ, et al. Definitions, methods, and applications in interpretable machine learning. *Proc Natl Acad Sci*. 2019;116(44):22071–80.
- [39] NIST. AI Risk Management Framework (AI RMF 1.0). 2023.
- [40] FUTURE-AI Consortium. International consensus guidelines for trustworthy AI in healthcare. 2025.
- [41] Liao QV, et al. Connecting Human-Centered and Technical Aspects of XAI. In: *AAAI*. 2023.
- [42] Belhajjame K, et al. PROV-O: The PROV Ontology. W3C Recommendation. 2013.
- [43] Wilkinson MD, et al. The FAIR Guiding Principles. *Sci Data*. 2016;3:160018.
- [44] Rudin C. Stop explaining black box models for high stakes decisions. *Nat Mach Intell*. 2019;1(5):206–15.
- [45] Ralinirina RE, Ralaivao JC, Ratovondrahona AJ, Mahatody T. A Formal Methodological Framework

for Auditing Robustness and Fidelity in Explainable AI: From Application to Trust Certification. In: Proceedings of the 13th International Conference on Artificial Intelligence (ICAI 2026), Eger, Hungary. 2026. doi:10.48550/arXiv.2608.23817

- [46] Rosa Elysabeth Ralinirina, et al. Benchmark Datasets for XAI Qualification. Zenodo. 2025. doi:10.5281/zenodo.20546322.