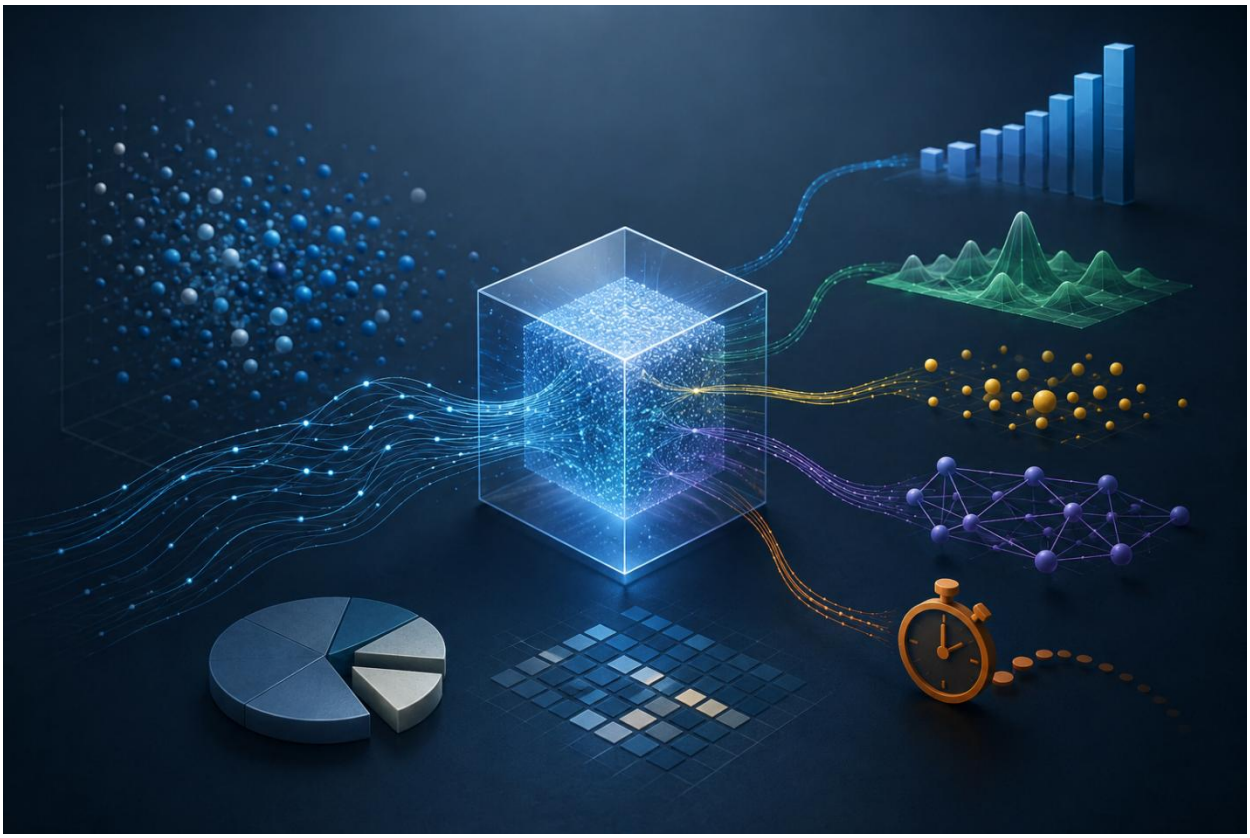
AÑO 4. No. 3
ISSN:2992-7978



A FRAMEWORK FOR RIGOROUS EVALUATION OF MODEL-AGNOSTIC EXPLAINABILITY METHODS: MULTI-METRIC STATISTICAL BENCHMARKING, OPERATIONAL PROTOCOL, AND REPRODUCIBILITY

Jonathan Herrera-Vasquez[1]
ORCID: 0000-0002-7149-6635
Miguel Herrero-Uceda[2]
ORDIC: 0009-0000-2715-6302

[1] Doctorando, Universidad Americana de Europa (UNADE). Correo: jonnabio@gmail.com
[2] Profesor-Investigador, Universidad Americana de Europa (UNADE). Correo: miguel.herrero.tutor@gmail.com

AÑO 4. No. 3
ISSN:2992-7978

Abstract

Evaluating explainability methods requires more than a single faithfulness proxy. We present a modular benchmarking framework centered on quantitative XAI quality metrics: fidelity, stability, sparsity, computational cost, and faithfulness gap, plus an explicit method for operating the framework end-to-end. On the UCI Adult benchmark, we use a staged evidence protocol: a calibration/reproducibility stage (EXP1) followed by a primary com-parative/robustness benchmark (EXP2); the current merged recovery snapshot contains 299 committed result artifacts (99.7% artifact coverage) plus a 30-row SHAP recovery batch, yielding 275 analyzable unique runs out of 300 planned cells (91.7%). Across complete model-size blocks (5 models, $N \in \{50, 100, 200\}$), Friedman tests indicate significant method differences for fidelity ($\chi^2 = 42.12$, $p = 3.78 \times 10^{-9}$), stability ($\chi^2 = 40.68$, $p = 7.65 \times 10^{-9}$), sparsity ($\chi^2 = 35.64$, $p = 8.92 \times 10^{-8}$), faithfulness gap ($\chi^2 = 45.00$, $p = 9.25 \times 10^{-10}$), and runtime ($\chi^2 = 30.44$, $p = 1.12 \times 10^{-6}$). SHAP leads on fidelity/stability, DiCE leads on sparsity, and LIME remains fastest overall. We release the framework, operation protocol, and artifacts with explicit data-quality caveats for reproducible benchmark use under a quantitative-only claim scope.

Keywords: Explainable AI, Benchmark Methodology, Reproducibility, Statistical Evaluation, Model-Agnostic Explanations

Resumen

La evaluación de los métodos de explicabilidad en inteligencia artificial (XAI) requiere considerar múltiples dimensiones más allá de la fidelidad. Este estudio propone un marco modular de evaluación comparativa basado en cinco métricas cuantitativas: fidelidad, estabilidad, dispersión, costo computacional y brecha de fidelidad, acompañado de un protocolo operativo reproducible. El marco se evaluó utilizando el conjunto de datos UCI Adult mediante dos etapas: calibración y reproducibilidad (EXP1), y comparación y robustez (EXP2). De las 300 ejecuciones planificadas, se obtuvieron 275 ejecuciones únicas analizables (91.7 %). Las pruebas de Friedman mostraron diferencias estadísticamente significativas entre los métodos evaluados en las cinco métricas. Los resultados indican que SHAP presenta el mejor desempeño en fidelidad y estabilidad, DiCE destaca en dispersión y LIME ofrece el menor costo computacional. Los hallazgos evidencian que no existe un método universalmente superior, sino diferentes compensaciones entre calidad explicativa, concisión y eficiencia computacional. El marco, el protocolo operativo y los artefactos resultantes se ponen a disposición para facilitar evaluaciones comparativas reproducibles bajo un alcance estrictamente cuantitativo.

Palabras clave: Inteligencia Artificial Explicable, Metodología De Evaluación Comparativa, Reproducibilidad, Evaluación Estadística, Explicaciones Agnósticas Al Modelo.

AÑO 4. No. 3
ISSN:2992-7978

INTRODUCTION

Problem Setting and Gap

Evaluation of post-hoc explainability methods remains methodologically fragmented. Recent reviews show heterogeneous metric choices and inconsistent evaluation objectives across domains, which weakens comparability and cumulative progress (Kadir et al., 2023; Pawlicki et al., 2024; Adadi and Berrada, 2018; Arrieta et al., 2020). Fidelity-centric reporting is still common even when stability, complexity, and computational burden materially affect practical usefulness, especially in high-stakes settings (Rudin, 2019).

Tooling has improved implementation-level reproducibility, including interfaces for broad explainer and metric coverage (Li et al., 2022; Hedström et al., 2023). However, tool avail-ability does not establish inferential comparability by itself. Benchmark claims still require explicit controls for split discipline, artifact qualification, analysis units, and multiplicity-aware statistical inference. Canonical model-agnostic explanation families (local surrogates, additive attributions, rule-based anchors, and counterfactual recourse) also rely on different assumptions, which complicates fair cross-method comparison (Ribeiro et al., 2016; Lundberg and Lee, 2017; Ribeiro et al., 2018; Mothilal et al., 2020).

Table 1

Novelty boundary relative to adjacent XAI benchmark resources.

Resource	Primary emphasis	Delta in this paper
OpenXAI (Agarwal et al., 2022)	Transparent evaluation of post-hoc explanations.	Adds an operation-level protocol that links execu-tion gates, artifact qualifi-cation, and claim eligibil-ity.
Quantus (Hedstrðm et al., 2023)	Toolkit support for many explanation quality met-rics.	Uses metrics inside a fixed inferential workflow with block-level tests, recovery overlays, and failure ac-counting.
BEExAI (Sithakoul et al., 2024)	Benchmarking infrastruc-ture for XAI comparison.	Emphasizes snapshot governance, deterministic analysis exports, and reproducibility profiling as first-class benchmark outputs.
This work	Model-agnostic tabular benchmark governance.	Provides FOM-7, a staged claim-control protocol connecting protocol freeze, controlled execution, integrity audit, statistical export, and bounded reporting.

Related Work and Novelty Delta

Recent benchmark efforts in graph explainability show that value depends on explicit dataset assumptions, ground-truth design, and transparent reporting constraints (Agarwal et al., 2023; Proszewska et al., 2025). Related findings on attribution consensus in Rashomon sets further indicate that explanation behavior can vary even under comparable predictive performance (Laberge et al., 2023). Faithfulness-oriented work also shows sensitivity to

AÑO 4. No. 3
ISSN:2992-7978

perturbation and protocol choices (Zheng et al., 2025).

The present study addresses a gap left by existing tools and benchmarks: a single, auditable protocol that links benchmark execution governance, artifact quality control, and inferential reporting. The methodological contribution is not a new explainer algorithm; it is a reproducible benchmark operation model that connects design decisions to claim eligibility.

Contributions and Scope

This paper contributes:

1. a model-agnostic benchmark framework spanning five model families and four explainer families under shared execution controls;
2. a staged evidence-generation protocol, spanning calibration/reproducibility (EXP1) and primary benchmarking (EXP2), with explicit artifact qualification and claim-role separa-tion;
3. a multi-metric inferential stack combining Friedman omnibus tests, Nemenyi localization, and paired Wilcoxon contrasts with multiplicity control;
4. an auditable **F**ramework **O**peration **M**ethod with seven gates (FOM-7) for protocol freeze, controlled execution, integrity audit, inferential export, and claim-traceable reporting.

FOM-7 emerged while consolidating early calibration-stage pilot cycles (EXP1), where protocol drift, malformed artifacts, and ad-hoc analysis-table harmonization repeatedly threatened claim comparability. The method was therefore formalized as a seven-gate operation protocol that makes claim eligibility auditable.

Scope is intentionally bounded to model-agnostic tabular post-hoc explanation bench-marking on Adult Income with quantitative metrics (fidelity, stability, sparsity, faithfulness gap, and cost). Overall, the current merged primary benchmark (EXP2) recovery snapshot combines 299 committed result artifacts with a 30-row SHAP recovery batch, yielding 275 analyzable unique runs out of 300 planned cells (91.7%). The remainder of this paper presents the methodology (Section 2), reports comparative and reproducibility evidence (Section 3), and concludes with implications, limitations, and next steps (Section 4).

METHODOLOGY

Study Scope and Evidence Cohorts

Benchmarking literature identifies two recurring barriers to reliable XAI comparison: non-standardized evaluation constructs and weak traceability between execution artifacts and reported claims. Recent frameworks and toolkits emphasize explicit property definitions, transparent benchmarking pipelines, and artifact-level reproducibility (Canha et al., 2025; Agarwal et al., 2022; Sithakoul et al., 2024; Hedström et al., 2023; Adadi and Berrada, 2018; Arrieta et al., 2020).

Following this guidance, this study separates evidence generation into staged cohorts with explicit claim roles. The separation is analogous to train/validation/test discipline, but applied to benchmark evidence rather than predictive model fitting. This design reduces leakage between calibration activities and confirmatory inference, a risk discussed in recent faithfulness evaluation work (Zheng et al., 2025), and it makes missingness/exclusion decisions auditable before inferential claims are finalized.

Research Questions, Factors, and Analysis Units

The confirmatory questions are:

AÑO 4. No. 3
ISSN:2992-7978

Table 2

Evidence cohorts and their role in this study's claims.

Cohort	Primary objective	Role in claims
Calibration sanity checks (EXP1 pilot)	Implementation checks before scaled runs.	Not used for confirmatory significance claims.
Reproducibility repeatability cohort (EXP1)	Repeated-seed variance profiling on core settings.	Supports reproducibility characterization (dispersion/CV).
Comparative screening cohort (EXP2 comparative)	Broad method-model coverage under fixed seeds.	Descriptive comparative context.
Primary robustness benchmark (EXP2 robustness)	Multi-seed, multi-sample-size benchmark runs.	Primary inferential cohort for global and paired tests.

Figure 1

Evidence qualification waterfall for primary robustness benchmark (EXP2) infer-ence.

Planned 300 → **Committed artifacts** 299 + **Recovery overlay** 30 → **Analyzable unique runs** 275 → **Friedman-ready blocks** 15/15

Overlay effect: 29 replacement MLP/SVM-SHAP runs + 1 newly recovered SHAP cell; residual unavailability = 25 empty Anchors/DiCE artifacts.

1. **RQ1 (global separation):** do explainer methods differ on the primary metrics under matched contexts?
2. **RQ2 (targeted pairwise contrast):** on matched cells, does SHAP differ from LIME on quality-cost trade-offs?
3. **RQ3 (reproducibility profile):** how much metric dispersion appears under repeated runs?

The primary robustness benchmark (EXP2) is a crossed design with model family g, explainer method k, seed s, and per-quadrant sampling intensity n:

$g \in \{\text{logreg}, \text{rf}, \text{xgb}, \text{svm}, \text{mlp}\}$, $k \in \{\text{shap}, \text{lime}, \text{anchors}, \text{dice}\}$, $s \in \{42, 123, 456, 789, 999\}$, $n \in \{50, 100, 200\}$.

Planned primary robustness benchmark (EXP2) size is 5 × 4 × 5 × 3 = 300 runs. In the current merged recovery snapshot, 299 result artifacts are committed under experiments/ exp2_scaled/results/; a 30-row SHAP recovery batch in outputs/batch_results.csv supersedes the mlp_shap/svm_shap slice, yielding 275 analyzable unique runs and leaving 25 unresolved cells after the overlay. The committed tree still has one missing SHAP artifact and 25 present-but-empty artifacts; the recovery overlay fills the missing SHAP cell for analysis.

Table 3

Primary benchmark (EXP2) evidence accounting for inferential qualification.

Qualification stage	Count	Share of planned runs
Planned robustness grid (5 × 4 × 5 × 3)	300	100.0%
Present artifacts	299	99.7%
Analyzable unique runs (committed + recovery overlay)	275	91.7%

AÑO 4. No. 3
ISSN:2992-7978

Unavailable cells after recovery overlay (25 empty)	25	8.3%
Complete (g, n) blocks used by Friedman tests	15/15	100.0% block completion

The inferential unit is not the instance row. For metric m, the run-level estimand is $y_{g,k,s,n,m}$ and the matched block is $b = (g, n)$. Tests are performed on run/block aggregates to avoid pseudo-replication.

Data Protocol and Leakage Controls

The benchmark uses Adult Income loaded through src/data_loading/adult.py with OpenML as primary source and UCI fallback. Cleaning applies whitespace normalization, missing-value canonicalization, target normalization, duplicate removal, and deterministic feature ordering. Deterministic preprocessing and artifact traceability follow reproducibility recommendations in open benchmarking toolkits (Agarwal et al., 2022; Hedström et al., 2023).

Leakage controls are fixed across all explainers:

1. stratified split before any preprocessing fit,
2. train-only preprocessor fitting,
3. shared persisted preprocessor across methods in each run,
4. explanation on the same transformed feature space used by the model.

Data split is

`train_test_split(test_size=0.2, stratify=y, random_state=seed)`.

For seed 42, the split is 39,032 train and 9,758 test rows. Each run samples TP/TN/FP/FN strata with target N per stratum (nominal run size $4N$, subject to stratum capacity).

Model and Explanation Protocol

The primary robustness benchmark (EXP2) evaluates model artifacts frozen during the calibration/reproducibility stage (EXP1); models are not retrained during the primary benchmark. Model families are logreg, rf, xgb, svm, and mlp, with fixed repository-tracked settings.

All methods are executed through ExplainerWrapper.explain_instance under a stan-dardized runtime constraint protocol (Cores=1, RAM=4GB, Timeout=300s) enforced by the

Table 4

Primary metric orientation for objective-consistent interpretation.

Metric	Unit	Direction	Operational interpretation
Fidelity	score	↑	Larger perturbation agreement indicates stronger local faithfulness under top-k masking.
Stability	cosine sim.	↑	Higher perturbation consistency indicates less explanation volatility.
Sparsity	score	↓	Lower value indicates more concise explanations under this implementation.
Faithfulness Gap	score	↑	Larger sensitivity gap indicates stronger alignment between importance ranking and predictive effect.
Cost	ms	↓	Lower runtime indicates higher practical deployability.

AÑO 4. No. 3
ISSN:2992-7978

infrastructure's resource guard to ensure comparable wall-clock duration reporting. Effective explainer conditions are:

1. SHAP: tree mode for RF and XGB; kernel mode for LogReg, SVM, and MLP; background sample count set to 50 (Lundberg and Lee, 2017).
2. LIME: num_samples=1000, num_features=10, kernel_width=3.0 (Ribeiro et al., 2016).
3. Anchors: AnchorTabular on transformed training data with effective threshold fixed at 0.95 (Ribeiro et al., 2018).
4. DiCE: random backend, opposite-class counterfactual target, importance from absolute original–counterfactual deltas (Mothilal et al., 2020).

Two implementation-bound caveats are explicit in claim interpretation: Anchors YAML threshold is not consumed at runtime (fixed 0.95), and DiCE YAML counterfactual count is not consumed (effective total_CFs=1).

Metrics and Statistical Inference

Primary metrics are fidelity, stability, sparsity, faithfulness gap, and cost (milliseconds). Aggregation is instance → run → block. EEU is not used for confirmatory inference in this draft. The multi-metric framing follows recommendations from recent XAI evaluation surveys and critical reviews (Kadir et al., 2023; Pawlicki et al., 2024; Adadi and Berrada, 2018; Arrieta et al., 2020).
Faithfulness and stability are computed as:

$$\Delta_k = |f(x) - f(x_{\text{mask-top-}k})|, \qquad \rho = \text{corr}(|w_i|, |f(x) - f(x_{-i})|), \tag{1}$$

$$x^{(t)} = x + E^{(t)},\ E^{(t)} \sim N(0, \sigma^2), \qquad S = \frac{2}{T(T-1)} \sum_{a<b} \cos\left(e^{(a)}, e^{(b)}\right). \tag{2}$$

Inferential workflow:

1. Friedman omnibus test per metric on complete (g, n) blocks.
2. Nemenyi post-hoc localization when Friedman is significant.
3. Two-sided paired Wilcoxon SHAP–LIME tests on matched (g, s, n) cells (45-cell primary set; 75-cell sensitivity set).

Multiplicity handling uses Nemenyi within each metric and Holm-Bonferroni across the five primary metrics for each inferential family (Friedman, Wilcoxon-45, Wilcoxon-75). Reported uncertainty/effect summaries include mean, standard deviation, CV, confidence intervals (where generated), Kendall's W, Cohen's d_z, and median paired differences.

Artifact qualification is deterministic: malformed or empty artifact-tree runs are excluded, empirical rerun summaries are accepted only from outputs/batch_results.csv for the mlp_shap/svm_shap recovery overlay, no synthetic reconstruction is used, and omnibus tests are restricted to block-complete contexts.

FOM-7 Origin, Operational Definition, and Validity Boundaries

FOM-7 denotes the **F**ramework **O**peration **M**ethod with seven sequential quality gates. It was derived from two inputs: (i) literature-grounded benchmark requirements for transparent, reproducible evaluation (Canha et al., 2025; Agarwal et al., 2022; Hedström et al., 2023), and (ii) failure modes observed in the calibration-stage pilot stream (EXP1; Table 2), especially protocol drift, artifact integrity failures, and non-equivalent analysis exports. To convert these risks into explicit controls, the benchmark workflow was restructured as the following stage-gated protocol:

AÑO 4. No. 3
ISSN:2992-7978

1. **Protocol freeze:** lock code/configuration versions, factors, and inferential plan before confirmatory runs.
2. **Controlled batch execution:** execute runs from declared manifests with fixed seeds and logged runtime context.
3. **Artifact integrity audit:** detect and exclude malformed, empty, or schema-incompatible outputs.
4. **Harmonization to analysis-ready tables:** standardize keys and metric fields into deterministic, merge-safe inferential tables.
5. **Inferential export:** generate deterministic omnibus and paired-test artifacts from qualified inputs only.
6. **Reproducibility profiling:** quantify run-to-run dispersion and coefficient-of-variation behavior on replicated settings.
7. **Claim-ready reporting:** permit inferential claims only when all prior gates are satisfied and traceable to source artifacts.

Progression requires gate satisfaction at each stage; failed gates downgrade affected outputs to descriptive-only status.

Figure 2

Compact operational flow of FOM-7 from protocol lock to claim-ready reporting.

FOM-7 flow: Freeze → Execute → Audit → Harmonize → Export → Profile → Report

The reproducibility package includes code, experiment configurations, model and prepro-cessor artifacts, per-run outputs, analysis exports, and figure scripts. Primary repository anchors used in this paper are:

- configs/experiments/exp2_scaled/manifest.yaml (declared primary benchmark (EXP2) run grid),
- experiments/exp2_scaled/results/ (qualified per-run outputs),
- outputs/paper_analysis/paper_comparison.csv (aggregated comparison export),
- scripts/generate_paper_a_figures.py (figure-generation entry point). Validity controls and residual risks are summarized as follows:

1. **Internal validity:** explicit missingness audit and block-complete inference. A standard-ized runtime constraint protocol (Cores=1, RAM=4GB, Timeout=300s) is configured in the experiment layer via the ResourceGuard utility to improve architectural comparability; residual risk from non-random coverage gaps.
2. **Construct validity:** metric definitions are operationalized in the src/metrics/ direc-tory. For technical traceability, conceptual definitions reported in Section 2.5 are an-chored to implementation classes: FidelityMetric, StabilityMetric, SparsityMetric, FaithfulnessMetric (gap field), and CostMetric.
3. **Statistical conclusion validity:** matched-design non-parametric tests with multiplicity control; residual risk from absent equivalence bands.
4. **External validity:** diversity across model families and seeds, but single tabular dataset scope.
5. **Implementation validity:** effective runtime parameter bindings are disclosed; harmo-nization changes may shift absolute levels.

Scope of claims: conclusions in Section 3 are comparative and bounded to this benchmark setting

AÑO 4. No. 3
ISSN:2992-7978

(Adult tabular task, declared models/explainers, and artifact-qualified primary benchmark (EXP2) subsets).

RESULTS

Overview of Evaluation Setup

This section reports results from two evidence streams: the primary robustness benchmark (EXP2) for global and paired method comparisons, and the reproducibility repeatability cohort (EXP1) for variance profiling. Compared explainers are SHAP, LIME, Anchors, and DiCE. Reported metrics are Fidelity, Stability, Sparsity, Faithfulness Gap, and Cost (ms). This multi-metric framing treats explanation quality as a vector of properties rather than a single scalar score (Kadir et al., 2023; Pawlicki et al., 2024; Arrieta et al., 2020). Metric definitions and the inferential workflow are specified in Section 2.5. "15 complete model-size blocks" denotes the block-level primary benchmark (EXP2) units used for global inference, and "45 matched configs" denotes the SHAP–LIME subset with shared (model, seed, *N*) cells on logreg/rf/xgb. Metric direction is defined in Table 4, and cohort qualification counts are reported in Table 3.

Global Trade-offs Across Methods (Quality vs Cost)

Table 5

Global method means (primary robustness benchmark (EXP2): 15 complete model-size blocks).

Method	Fidelity	Stability	Sparsity	Faithfulness Gap	Time (ms)
SHAP	0.8081	0.7320	0.2264	0.3796	11708.26
LIME	0.5602	0.0144	0.0846	0.3342	3660.68
Anchors	0.3886	0.0429	0.0851	0.2361	32463.44
DiCE	0.1701	0.3611	0.0166	0.0946	27940.35

SHAP has the highest Fidelity (0.8081), highest Stability (0.7320), and highest Faithful-ness Gap (0.3796). DiCE has the lowest Sparsity value (0.0166), while LIME has the lowest Cost (3660.68 ms). In the merged recovery snapshot, SHAP's mean Cost is 11708.26 ms, remaining above LIME but below Anchors and DiCE.

The reported means show an explicit quality–cost frontier: SHAP occupies the strongest quality region on Fidelity/Stability/Faithfulness Gap with a substantially reduced mean runtime (11708.26 ms) relative to earlier artifact-only snapshots, whereas LIME occupies the lowest-cost region with lower Fidelity and Stability. DiCE provides the lowest Sparsity value with lower Fidelity and Faithfulness Gap. This trade-off pattern is consistent with mechanistic differences between local surrogate and additive-attribution families (Ribeiro et al., 2016; Lundberg and Lee, 2017).

Global Statistical Evidence Across Methods

Table 6

Friedman tests across methods.

AÑO 4. No. 3
ISSN:2992-7978

Metric	Statistic	p-value	Significant
Fidelity	42.12	3.78e-09	Yes
Stability	40.68	7.65e-09	Yes
Sparsity	35.64	8.92e-08	Yes
Faithfulness Gap	45.00	9.25e-10	Yes
Cost	30.44	1.12e-06	Yes

For every metric, the Friedman test rejects the null of equal performance across methods. This establishes global between-method differences, but does not identify pairwise direction without post-hoc localization. Post-hoc localization uses Nemenyi pairwise comparisons within each metric, with Holm-Bonferroni adjustment across the five primary metrics. The corresponding effect sizes indicate medium-to-very-large between-method separation across all five metrics.

Table 7

Friedman effect sizes (Kendall's $W = \chi^2/[N(k-1)]$; $N = 15$, $k = 4$).

Metric	χ^2	Kendall's W
Fidelity	42.12	0.936
Stability	40.68	0.904
Sparsity	35.64	0.792
Faithfulness Gap	45.00	1.000
Cost	30.44	0.676

Targeted Pairwise Comparison (SHAP vs LIME on Shared Families)

Table 8

Paired Wilcoxon comparison on 45 matched configurations (logreg/rf/cgb)

Metric	SHAP Mean	LIME Mean	Wilcoxon p-value
Fidelity	0.8112	0.5556	5.68e-14
Stability	0.8013	0.0154	5.68e-14
Sparsity	0.3156	0.0845	5.68e-14
Faithfulness Gap	0.3924	0.3408	5.68e-14
Cost (ms)	1258.72	210.45	2.29e-06

Within this matched subset, SHAP is higher on Fidelity, Stability, and Faithfulness Gap, while LIME has lower Cost and lower Sparsity value. The reported p-values indicate non-zero paired differences for all listed metrics. Practically, the quality gains for SHAP are accompanied by a runtime

AÑO 4. No. 3
ISSN:2992-7978

penalty relative to LIME under shared cells.

This comparison is restricted to logreg/rf/xgb matched configurations and is not directly generalizable to non-matched families or to methods outside the SHAP–LIME pair, including rule-based and counterfactual families (Ribeiro et al., 2018; Mothilal et al., 2020).

- **Reproducibility and Variance Profile**

The reproducibility repeatability cohort (EXP1) indicates lower dispersion for quality-oriented metrics and higher dispersion for runtime behavior. Quality metrics remain relatively stable across repeated seeds, while Cost varies more strongly for SHAP variants, consistent with recent calls for variance-aware benchmark reporting (Agarwal et al., 2022; Hedström et al., 2023).

- Metric stability (quality): Fidelity CV ≤ 0.0634, Stability CV ≤ 0.0826, Sparsity CV ≤ 0.0441, Faithfulness-gap CV ≤ 0.0358.
- Runtime instability (cost): Cost CV reaches 0.225 for SHAP variants.
- Decision implication: quality rankings are more repeatable than runtime behavior under the reported repeated-run protocol.

- **Synthesis and Implications for Practice**

- When the objective is maximizing Fidelity/Stability/Faithfulness Gap and compute budget is permissive, SHAP is the supported choice in this benchmark.
- When compute or latency is constrained, LIME is the supported low-cost option, with lower Fidelity and Stability than SHAP.
- When conciseness (Sparsity) is the primary objective, DiCE provides the lowest reported Sparsity value, with lower Fidelity and Faithfulness Gap.
- Methods are not statistically interchangeable: Friedman tests reject the null of equal performance across methods for all five primary metrics.
- On the shared SHAP–LIME subset, quality improvements for SHAP are paired with higher runtime cost.

Interpretation limits:

- Anchors stability values near zero should be interpreted with implementation caveats: effective runtime behavior uses fixed precision threshold 0.95 and binary rule-membership importance vectors, which can reduce perturbation-based similarity under this protocol.
- Method rankings are currently validated only on one tabular benchmark (Adult); external generalization requires replication on additional datasets and modalities before cross-domain claims are made.

Formal Problem Framing and Scope of XAI Evaluation

Recent work argues that many XAI evaluations are method-first rather than problem-first, which can obscure what a given metric actually validates (Haufe et al., 2026). Following this critique, we treat explanation quality as purpose-conditioned: an explanation is evaluated relative to a clearly specified stakeholder question, not as a universal property of a method. In this study, we therefore distinguish between correctness-oriented criteria and secondary quality indicators. Correctness-oriented criteria address whether an explanation supports the intended inferential use under stated assumptions; secondary indicators include robustness, stability, sparsity, and computational efficiency, which are informative but not sufficient on their own to establish explanation validity (Doshi-Velez

AÑO 4. No. 3
ISSN:2992-7978

and Kim, 2017; Hedström et al., 2023).

A central risk identified in recent theory is that attribution methods may assign high importance to suppressor variables, that is, variables that improve prediction but are not associated with the target in the intended explanatory sense (Wilming et al., 2022, 2023; Haufe et al., 2026). This possibility motivates conservative interpretation of attribution rankings, especially in dependent-feature settings. Accordingly, we report method performance as
comparative evidence within the benchmark design, while avoiding the stronger claim that higher benchmark scores necessarily imply explanation correctness for all use cases. Where possible, we interpret results as compatibility with specific explanatory goals rather than definitive proof of faithful explanation.

Operationally, we adopt a requirements-to-metric traceability approach: each evaluation objective is tied to an intended use, an explicit assumption set, and a statistical test. This framing is consistent with formalization-oriented recommendations in recent XAI governance and assessment literature (Sokol and Flach, 2020; DIN, 2023; Zicari et al., 2021). Under this design, our benchmark is intended to support transparent comparison and hypothesis generation about method behavior, while acknowledging that full correctness claims may require additional theoretical analysis or controlled ground-truth experiments (Haufe et al., 2026).

Responsible Use and Benchmark Boundaries

This benchmark should not be read as evidence that post-hoc explanations are appropriate for all high-stakes decision settings. The Adult Income task contains demographic and socioeconomic variables, and benchmark rankings over this dataset can reflect properties of the task, preprocessing choices, model families, and metric implementations rather than universal properties of an explainer family. In particular, higher scores on the reported quantitative metrics do not imply fairness, causal validity, user trust, or suitability for deployment in eligibility, employment, credit, or other consequential settings (Rudin, 2019). The intended use of the released artifacts is methodological: reproducing the reported benchmark, auditing the effect of artifact qualification and statistical filtering, and extending the protocol to additional datasets or explanation families. Claims about deployment utility, human interpretability, or cross-domain method superiority require additional validation
beyond the quantitative-only scope of this paper.

Code and Artifact Availability

The current refreshed analysis snapshot is publicly available at https://github.com/jonnabio/xai-eval-framework and should be archived under a new versioned release and
DOI before submission. A prior submission snapshot for release tag paper-a-submission-2026-03-28 has version-specific DOI 10.5281/zenodo.19297724, but that prior DOI does not by itself identify the refreshed primary benchmark (EXP2) evidence cut reported here. Materials required to reproduce the Paper A analyses include experiments/exp2_scaled/results/, outputs/batch_results.csv, outputs/analysis/paper_a_exp2_stats/, scripts/run_exp2_statistical_analysis.py, scripts/generate_paper_a_figures.py, and the manuscript sources under docs/reports/paper_a/. A path-level artifact index is maintained in docs/reports/paper_a/paper_a_artifact_index.md. The repository is released under
the MIT License.

AÑO 4. No. 3
ISSN:2992-7978

CONCLUSION

The reported metrics indicate a multi-objective frontier rather than a single best method: SHAP is strongest on the reported quality-oriented metrics, LIME is strongest on runtime efficiency, and DiCE is strongest on the reported sparsity criterion. Global tests reject the null of equal method performance across the primary metrics, and paired SHAP–LIME analysis indicates a quality–cost trade-off on shared model families. This reinforces multi-objective interpretation guidance in current XAI evaluation literature (Pawlicki et al., 2024; Adadi and Berrada, 2018; Arrieta et al., 2020).

Methodologically, the current paper contributes FOM-7 (Framework Operation Method, seven-stage gating) as an auditable benchmark operation protocol that connects protocol freezing, controlled execution, integrity audit, deterministic inference export, and claim-traceable reporting. This contribution is operational and reproducibility-focused rather than algorithmic.

- **Takeaway:** method choice should be objective-driven (quality, sparsity, or runtime), not based on a single aggregated score.
- **Takeaway:** statistical evidence supports method differentiation at the omnibus level; pairwise direction must be interpreted within matched-design limits.
- **Limitation:** Current primary benchmark (EXP2) availability gaps remain concentrated in Anchors and DiCE; after integrating the 30-row SHAP recovery batch, 25 unavailable cells still qualify confidence in full-grid frontier stability. The reporting caveats document the current merged-snapshot diagnostic.
- **Next steps:** Commit the recovered svm_shap_s456_n200 per-run artifact when the active claim completes, diagnose the remaining empty Anchors/DiCE artifacts, and maintain the versioned reproducibility bundle (code, configs, run outputs, inferential exports, and figure scripts) synchronized to the current 275 analyzable-run snapshot.
- **Next steps:** replicate this quantitative protocol on additional tabular datasets to test ranking stability under dataset shift.

Acknowledgments and Disclosure of Funding

This work received no external funding. The study was self-funded by the authors, and the authors report no competing interests. This draft was prepared from repository artifacts dated April 2026 and should be paired with a refreshed archival DOI before submission.

REFERENCES

Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160.

Chirag Agarwal, Dan Ley, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. 2022. OpenXAI: Towards a Transparent Evaluation of Post hoc Model Explanations. *arXiv preprint arXiv:2206.11104*.

Chirag Agarwal, Owen Queen, Himabindu Lakkaraju, and Marinka Zitnik. 2023. Evaluating

AÑO 4. No. 3
ISSN:2992-7978

explainability for graph neural networks. *Scientific Data*, 10:144.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelli-gence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115.

Dulce Canha, Sylvain Kubler, Kary Främling, and Guy Fagherazzi. 2025. A Functionally-Grounded Benchmark Framework for XAI Methods: Insights and Foundations from a Systematic Literature Review. *ACM Computing Surveys*, 57(12):1–40.

Anna Hedström, Leander Weber, Dilyara Bareeva, Daniel Krakowczyk, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M.-C. Höhne. 2023. Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond. *Journal of Machine Learning Research*, 24:1–11.

Md Abdul Kadir, Amir Mosavi, and Daniel Sonntag. 2023. Evaluation Metrics for XAI: A Review, Taxonomy, and Practical Applications. In *Proceedings of the 2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES)*, pages 111–124, Nairobi, Kenya.

Gabriel Laberge, Yann Pequignot, Alexandre Mathieu, Foutse Khomh, and Mario Marchand. 2023. Partial Order in Chaos: Consensus on Feature Attributions in the Rashomon Set. *Journal of Machine Learning Research*, 24:1–50.

Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predic-tions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774.

Xuhong Li, Haoyi Xiong, Xingjian Li, Xuanyu Wu, Zeyu Chen, and Dejing Dou. 2022. InterpretDL: Explaining Deep Models in PaddlePaddle. *Journal of Machine Learning Research*, 23:1–6.

Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 607–617.

Marek Pawlicki, Aleksandra Pawlicka, Federica Uccello, Sebastian Szelest, Salvatore D'An-tonio, Rafał Kozik, and Michał Choraś. 2024. Evaluating the necessity of the multiple metrics for assessing explainable AI: A critical examination. *Neurocomputing*, 602:128282.

Samuel Sithakoul, Sara Meftah, and Clément Feutry. 2024. BEExAI: Benchmark to Evaluate Explainable AI. *arXiv preprint arXiv:2406.00596*.

Magdalena Proszewska, Tomasz Danel, and Dawid Rymarczyk. 2025. B-XAIC Dataset: Benchmarking Explainable AI for Graph Neural Networks Using Chemical Data. *arXiv preprint arXiv:2505.22252*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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*, pages 1135–1144.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Xu Zheng, Farhad Shirani, Zhuomin Chen, Chaohao Lin, Wei Cheng, Wenbo Guo, and Dongsheng Luo. 2025. F-FIDELITY: A Robust Framework for Faithfulness Evaluation of Explainable AI. In *International Conference on Learning Representations (ICLR)*.

AÑO 4. No. 3
ISSN:2992-7978

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity Checks for Saliency Maps. In *Advances in Neural Information Processing Systems*, volume 31, pages 9525–9536.

Gagan Bansal, Tongshuang Wu, Joel Vaughan, and Hanna Wallach. 2021. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 454–464.

DIN Deutsches Institut für Normung e. V. 2023. DIN SPEC 92001-3:2023-04. Artificial Intelligence—Life Cycle Processes and Quality Requirements—Part 3: Explainability.

Finale Doshi-Velez and Been Kim. 2017. Towards a Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608*.

Raymond Fok and Daniel S. Weld. 2023. In Search of Verifiability: Explanations Rarely Enable Complementary Performance in AI-Advised Decision Making. *AI Magazine*, 45(3):317–332.

Stefan Haufe, Rick Wilming, Benedict Clark, Rustam Zhumagambetov, Ahcène Boubekki, Jörg Martin, and Danny Panknin. 2026. Explainable AI Needs Formalization. *npj Artificial Intelligence*, 2:42. https://doi.org/10.1038/s44387-026-00095-1.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. A Benchmark for Interpretability Methods in Deep Neural Networks. In *Advances in Neural Information Processing Systems*, volume 32, pages 9737–9748.

Alon Jacovi and Yoav Goldberg. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205.

Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. 2022. A Consistent and Efficient Evaluation Strategy for Attribution Methods. In *Proceedings of the 39th International Conference on Machine Learning*, pages 18770–18795.

Kacper Sokol and Peter Flach. 2020. Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pages 56–67.

Rick Wilming, Christian Budding, Klaus-Robert Müller, and Stefan Haufe. 2022. Scrutinizing XAI Using Linear Ground-Truth Data with Suppressor Variables. *Machine Learning*, pages 1–21.

Rick Wilming, Lisa Kieslich, Benedict Clark, and Stefan Haufe. 2023. Theoretical Behavior of XAI Methods in the Presence of Suppressor Variables. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 37091–37107.

Roberto V. Zicari et al. 2021. Z-Inspection®: A Process to Assess Trustworthy AI. *IEEE Transactions on Technology and Society*, 2(2):83–97.