Journal of Science, Innovation & Social Impact

2026 2nd International Conference on Intelligent Computing and Automated Systems (ICAS 2026)

Article

Comparative Evaluation of Post-Hoc Feature Attribution Methods on Tabular Financial Data: Faithfulness, Stability, and Computational Efficiency

Pengyuan Xiao [1,*] and Xuanyi Fu [2]

[1] Computer Science, Zhejiang University, Hangzhou, China
[2] M.S.E. in Computer Science, Johns Hopkins University, Baltimore, MD, USA
* Correspondence: Pengyuan Xiao, Computer Science, Zhejiang University, Hangzhou, China

Abstract: The deployment of machine learning in credit scoring and fraud detection has intensified regulatory and societal demand for transparent decision-making. Post-hoc feature attribution methods such as SHAP, LIME, Integrated Gradients, and Anchors promise to explain individual predictions, yet their comparative reliability on financial tabular data remains insufficiently characterized. This study conducts a controlled empirical evaluation of four prominent attribution methods across four public financial datasets spanning credit scoring and transaction fraud detection. Three classifiers---XGBoost, Random Forest, and Multilayer Perceptron---serve as the underlying predictive functions. Explanation quality is quantified along three axes: faithfulness measured by Prediction Gap on Important features and infidelity, stability measured by max-sensitivity, and computational efficiency measured by wall-clock time per explanation. Results indicate that TreeSHAP achieves the highest faithfulness and lowest sensitivity on tree-based classifiers, while Integrated Gradients attains competitive faithfulness on neural networks. LIME exhibits the largest variance across repeated runs, raising concerns for regulatory settings that require reproducible explanations. Anchors produce the sparsest explanations at the cost of reduced faithfulness. No single method dominates all evaluation criteria simultaneously, corroborating recent theoretical predictions of an inherent trade-off among explanation desiderata. These findings provide practitioners and regulators with empirically grounded guidance for selecting attribution methods in financial applications.

Keywords: explainable artificial intelligence; feature attribution; credit scoring; faithfulness evaluation

Received: 09 March 2026
Revised: 20 April 2026
Accepted: 01 May 2026
Published: 06 May 2026

1. Introduction

1.1. Background and Motivation

Machine learning algorithms increasingly underpin high-stakes financial decisions ranging from consumer credit approval to real-time transaction monitoring. Regulatory frameworks in multiple jurisdictions now mandate that automated decisions affecting individuals be accompanied by meaningful explanations. Bücker et al [1]. demonstrated that transparency, auditability, and explainability constitute three distinct requirements for credit scoring, each imposing specific technical constraints on the deployed solution. The European Union AI Act, which classifies credit scoring as a high-risk application, further elevates explainability from a desirable property to a legal obligation. Financial institutions that adopt opaque gradient-boosted ensembles or deep networks must

reconcile the predictive advantage of these algorithms with the interpretability expected by regulators, auditors, and affected consumers.

Post-hoc feature attribution methods have emerged as the predominant technical response to these requirements. SHAP assigns each feature a Shapley value derived from cooperative game theory, providing a theoretically grounded decomposition of individual predictions [2]. LIME constructs local linear surrogates by perturbing input instances and fitting interpretable approximations in a neighborhood defined by an exponential kernel [3]. These two methods, along with gradient-based alternatives and rule-extraction approaches, form the current practitioner toolkit for explaining financial machine learning predictions. Their adoption is widespread, yet their relative merits on financial tabular data have not been subjected to rigorous side-by-side evaluation using standardized metrics.

1.2. Research Gap and Contributions

Despite widespread adoption, the reliability of these methods on financial tabular data has received limited systematic scrutiny. Krishna et al. formalized the disagreement problem, documenting that state-of-the-art explanation methods frequently assign conflicting importance rankings to the same features---a finding with serious implications for domains where explanation consistency carries legal weight [4]. Through practitioner interviews and empirical analysis, they demonstrated that data scientists rely on ad hoc heuristics to resolve these disagreements, underscoring the need for objective comparative evaluation. While several benchmarking efforts have addressed image and text modalities, the unique characteristics of financial tabular data---mixed feature types, strong inter-feature correlations, extreme class imbalance, and domain-specific regulatory constraints---warrant dedicated investigation.

1.3. Research Questions

This study addresses three specific questions. The first examines which attribution method achieves the highest faithfulness to the underlying classifier on financial tabular datasets of varying scale and imbalance. The second investigates how explanation stability differs across methods when input instances are subject to small perturbations, simulating the sensitivity analyses common in financial model validation. The third quantifies the computational cost of each method and assesses whether efficiency-faithfulness trade-offs favor different methods at different dataset scales. By answering these questions with controlled experiments on public data, this study aims to provide actionable guidance for practitioners navigating the increasingly complex landscape of explainable financial AI.

1.4. Paper Organization

The remainder of this paper proceeds as follows. Section 2 reviews related work on explanation methods, evaluation metrics, and financial applications. Section 3 details the experimental setup including datasets, classifiers, explanation methods, and evaluation protocol. Section 4 presents quantitative results with statistical analysis. Section 5 discusses implications, limitations, and directions for future research.

2. Related Work

2.1. Post-Hoc Feature Attribution Methods

2.1.1. Perturbation-Based Approaches

LIME generates explanations by sampling perturbed instances around a target prediction, weighting them by proximity through an exponential kernel, and fitting a sparse linear surrogate. While its model-agnostic nature makes it broadly applicable, Alvarez-Melis and Jaakkola identified instability as a fundamental limitation: repeated LIME runs on the same instance can yield substantially different explanations due to stochastic sampling of the perturbation neighborhood [5]. Anchors extend the perturbation paradigm by producing IF-THEN rules that hold above a user-specified

precision threshold, offering discrete and highly interpretable outputs at the cost of lower coverage on continuous features.

2.1.2. Gradient-Based and Game-Theoretic Approaches

Sundararajan et al. proposed Integrated Gradients, which computes attributions by accumulating gradients along a straight-line path from a baseline input to the actual input [6]. The method satisfies two key axioms---sensitivity and implementation invariance---providing formal guarantees that attribution methods based on single-point gradients cannot offer. SHAP unifies multiple explanation paradigms under the Shapley value framework, with TreeSHAP exploiting tree structure for exact polynomial-time computation and KernelSHAP offering model-agnostic approximation through weighted kernel regression. Adebayo et al. introduced sanity checks demonstrating that some gradient-based saliency methods produce attributions insensitive to both model parameters and training data, establishing minimum validity criteria [7]. Both Integrated Gradients and SHAP pass these sanity checks, placing them on firmer theoretical ground than methods such as Guided Backpropagation that fail the model randomization test.

2.2. Evaluation Metrics and Benchmarks

Quantitative evaluation of explanations has progressed from ad hoc assessment to systematic benchmarking. Yeh et al. formalized infidelity---measuring the expected squared discrepancy between attribution-predicted and actual perturbation effects---and max-sensitivity, quantifying explanation change within a local input neighborhood [8]. Hooker et al. proposed ROAR, which retrains classifiers after removing attributed features to measure faithfulness, and found that many popular methods performed no better than random feature removal [9]. Agarwal et al. introduced OpenXAI, a benchmark with twenty-two evaluation metrics spanning faithfulness, stability, and fairness, applied to tabular datasets including German Credit [10]. Han et al. proved a no-free-lunch result for post-hoc explanations: no single method achieves optimal local faithfulness across all neighborhoods, providing theoretical support for the comparative approach taken in this work [11].

2.3. Explainability in Financial Applications

Financial applications present distinctive challenges for explanation methods that go beyond standard machine learning settings. Feature correlations in credit data---between income and debt-to-income ratio, between credit utilization and payment history---violate the independence assumptions underlying marginal perturbation schemes in both LIME and KernelSHAP. Gramegna and Giudici compared SHAP and LIME on credit risk data and found that SHAP-derived feature weights provided more stable discriminative signals for downstream clustering tasks, attributing this advantage to the consistency properties of the Shapley value decomposition [12]. Bracke et al. applied Quantitative Input Influence with Shapley values to mortgage default prediction at the Bank of England, identifying loan-to-value ratio and current interest rate as primary default drivers---results consistent with established economic theory, lending credibility to the attribution method [13]. Their work demonstrated that explanation outputs can be validated against domain knowledge, a strategy this study adopts when interpreting the feature rankings produced by each method.

3. Experimental Setup

3.1. Datasets and Preprocessing

Four publicly available financial datasets are selected to span a range of sample sizes, feature dimensionalities, and class imbalance ratios. Table 1 summarizes their characteristics. The selection criteria prioritize datasets with verified provenance, documented feature semantics, and sufficient adoption in prior explainability research to enable comparison with published baselines.

Table 1. Dataset Characteristics

Dataset	Samples	Features	Feature Types	Positive Rate	Source
German Credit	1,000	20	13 categorical, 7 numerical	30.0%	UCI ML Repository
Taiwan Credit	30,000	23	Mixed integer and real	22.1%	UCI ML Repository
FICO HELOC	10,459	23	Numerical and ordinal	52.2%	FICO xML Challenge
Credit Card Fraud (CCF)	284,807	30	28 PCA + Time + Amount	0.172%	ULB / Kaggle

Data sources: German Credit and Taiwan Credit from the UCI Machine Learning Repository (CC BY 4.0); FICO HELOC from the FICO Explainable ML Challenge (free with registration); Credit Card Fraud from Kaggle (ODbL license).

German Credit contains 1,000 loan applicants described by 20 mixed-type attributes with a 30% default rate, making it the smallest and most class-balanced dataset in this evaluation. Taiwan Credit comprises 30,000 credit card holders with 23 features capturing six months of repayment history and billing information, offering a medium-scale credit scoring benchmark. FICO HELOC provides 10,459 Home Equity Line of Credit applications with 23 numerical predictors and near-balanced labels; this dataset was explicitly designed for explainability research by FICO as part of their Explainable Machine Learning Challenge. Credit Card Fraud contains 284,807 European card transactions collected over two days, with only 492 fraudulent cases (0.172%), representing extreme class imbalance that tests explanation robustness under severe distributional skew.

Preprocessing follows a standardized pipeline applied uniformly across all experiments. Categorical features are one-hot encoded for the Multilayer Perceptron and left as native categories for tree-based classifiers. Numerical features are standardized to zero mean and unit variance for the Multilayer Perceptron; no scaling is applied for tree-based classifiers, which are invariant to monotone feature transformations. Missing values in FICO HELOC (coded as −9, −8, −7 in the original data) are imputed with the training-fold median, following the preprocessing conventions adopted by Misheva et al. in their credit risk benchmarks. Stratified five-fold cross-validation is used throughout, with explanation evaluation performed exclusively on test-fold instances to prevent information leakage [14].

3.2. Explanation Methods and Configuration

3.2.1. Method Selection and Implementation

Four attribution methods are evaluated on tree-based classifiers. TreeSHAP computes exact Shapley values by exploiting the internal structure of tree ensembles, producing deterministic explanations with no sampling variance. KernelSHAP approximates Shapley values through weighted linear regression on randomly sampled feature coalitions, configured with 1,000 background samples per explanation to balance approximation quality and computational cost. LIME generates local linear surrogates using 5,000 perturbed instances per explanation, with an exponential kernel whose bandwidth is set to 0.75 times the square root of the feature count---a heuristic that adapts to varying feature dimensionalities across datasets. Anchors produces rule-based explanations using beam search with a precision threshold of 0.95 and a confidence parameter of 0.95, following the default configuration recommended by the original authors.

On the Multilayer Perceptron, KernelSHAP, LIME, and Integrated Gradients are evaluated. Integrated Gradients accumulates gradients along 200 interpolation steps from

a zero baseline to the input instance, with the step count chosen to ensure convergence of the integral approximation while keeping computation tractable.

3.2.2. Hyperparameter Settings

Three classifiers serve as the predictive functions to be explained, representing the dominant architectures in financial machine learning practice. XGBoost is configured with 200 trees, maximum depth 6, learning rate 0.1, and column subsampling ratio 0.8. Random Forest uses 200 trees with unrestricted depth and minimum samples per leaf set to 5. The Multilayer Perceptron contains two hidden layers of 128 and 64 units with ReLU activations, trained for 100 epochs with Adam optimization (learning rate 0.001) and early stopping on validation loss with a patience of 10 epochs. Table 2 reports the predictive performance of each classifier across all four datasets.

Table 2. Base Classifier Performance (AUC-ROC, 5-Fold Cross-Validation)

Classifier	German	Taiwan	FICO	CCF
XGBoost	0.782	0.785	0.798	0.977
Random Forest	0.771	0.774	0.784	0.970
MLP	0.754	0.766	0.771	0.973

All values are means over 5 stratified folds. Standard deviations range from 0.008 to 0.023 across configurations.

XGBoost achieves the highest AUC-ROC on all four datasets, consistent with the well-documented advantage of gradient-boosted trees on heterogeneous tabular data. The performance gap between XGBoost and MLP is largest on German Credit (0.028 AUC difference) where the small sample size limits the effectiveness of neural network training. All subsequent explanation experiments use XGBoost as the primary classifier, with MLP serving as the secondary classifier for Integrated Gradients evaluation.

3.3. Evaluation Protocol

3.3.1. Faithfulness Metrics

Faithfulness is assessed using two complementary metrics derived from the formal definitions established in the evaluation literature. Prediction Gap on Important features (PGI) measures the mean absolute change in predicted probability when the top-K most important features (K = 5) are replaced with values drawn from the training marginal distribution. Higher PGI indicates that the attribution method successfully identifies features whose removal most disrupts the prediction. The evaluation follows the Noisy Linear Imputation strategy proposed by Rong et al. to mitigate the out-of-distribution artifacts that arise when features are simply zeroed or mean-imputed, which can inflate faithfulness estimates by introducing implausible feature combinations [15].

Infidelity quantifies the expected squared difference between the dot product of attributions with a perturbation vector and the actual prediction change induced by that perturbation. Lower infidelity indicates closer alignment between attributed importance and true prediction sensitivity. Gaussian perturbations with standard deviation 0.1 are applied, and expectations are estimated over 500 Monte Carlo samples per test instance. This sample size achieves coefficient-of-variation below 0.05 on the infidelity estimator across all dataset-method configurations.

3.3.2. Stability and Efficiency Metrics

Stability is measured by max-sensitivity, defined as the maximum change in explanation norm within an ε-ball around the input instance. For each test instance, 50 neighbors are sampled uniformly within an L2-norm ball of radius $\varepsilon = 0.01$, and the maximum L2 distance between the original explanation and each perturbed explanation is recorded. Lower max-sensitivity indicates more stable explanations that are robust to small input variations---a property of particular importance in financial model validation, where sensitivity analysis is a standard component of model risk management. The Quantus toolkit provides the implementation for all stability computations, ensuring

reproducibility and consistency with the metric definitions used in recent benchmarking studies [16].

Sparsity is quantified by the Gini coefficient of absolute attribution values, where higher values indicate that importance is concentrated on fewer features. A Gini coefficient of 1.0 would indicate all importance assigned to a single feature, while 0.0 would indicate perfectly uniform distribution. Computational efficiency is measured as wall-clock time in seconds per explanation, averaged over 200 randomly sampled test instances, executed on a single CPU core (Intel Xeon E5-2680 v4 at 2.40 GHz) to ensure comparability across methods.

4. Results and Analysis

4.1. Faithfulness Evaluation

4.1.1. Feature Removal Perturbation Results

Table 3 reports PGI and infidelity scores for all four attribution methods on XGBoost across the four datasets. TreeSHAP achieves the highest PGI on every dataset, with scores ranging from 0.138 on Taiwan Credit to 0.168 on Credit Card Fraud. KernelSHAP follows consistently, trailing TreeSHAP by 0.011 to 0.016 PGI points across datasets. LIME ranks third on all datasets, while Anchors produces the lowest PGI scores throughout.

Table 3. Faithfulness Metrics on XGBoost (Mean ± Std, 5-Fold CV)

Method	German PGI (↑)	Taiwan PGI (↑)	FICO PGI (↑)	CCF PGI (↑)
TreeSHAP	0.142 ± 0.011	0.138 ± 0.008	0.155 ± 0.009	0.168 ± 0.006
KernelSHAP	0.131 ± 0.014	0.125 ± 0.010	0.139 ± 0.012	0.152 ± 0.008
LIME	0.108 ± 0.021	0.102 ± 0.018	0.114 ± 0.019	0.121 ± 0.015
Anchors	0.087 ± 0.016	0.082 ± 0.014	0.093 ± 0.015	0.095 ± 0.012

PGI computed with K = 5 features replaced via Noisy Linear Imputation. Higher PGI indicates stronger faithfulness.

The gap between TreeSHAP and LIME widens from 0.034 on German Credit to 0.047 on Credit Card Fraud, suggesting that the faithfulness advantage of Shapley-based methods grows on larger and more imbalanced datasets where the perturbation neighborhood structure becomes more complex. LIME's standard deviations are approximately twice those of TreeSHAP across all datasets (0.015--0.021 versus 0.006--0.011), reflecting the additional variance introduced by stochastic neighborhood sampling. Anchors lag behind continuous attribution methods by a substantial margin, likely because their binary rule format cannot capture fine-grained feature contributions and must discretize inherently continuous decision boundaries.

4.1.2. Infidelity Score Analysis

Infidelity scores reinforce the PGI ranking with a consistent method ordering across all four datasets. TreeSHAP achieves the lowest infidelity (0.016--0.024), indicating tight alignment between its attributions and actual prediction sensitivity under Gaussian perturbation. KernelSHAP infidelity values (0.023--0.031) are approximately 30% higher than TreeSHAP, a gap attributable to the sampling approximation of the Shapley value computation. LIME infidelity is approximately 2.4 times that of TreeSHAP across datasets. Anchors exhibit the highest infidelity (0.075--0.089), reflecting the fundamental information loss inherent in discretizing continuous attributions into binary rules (As shown in Figure 1).

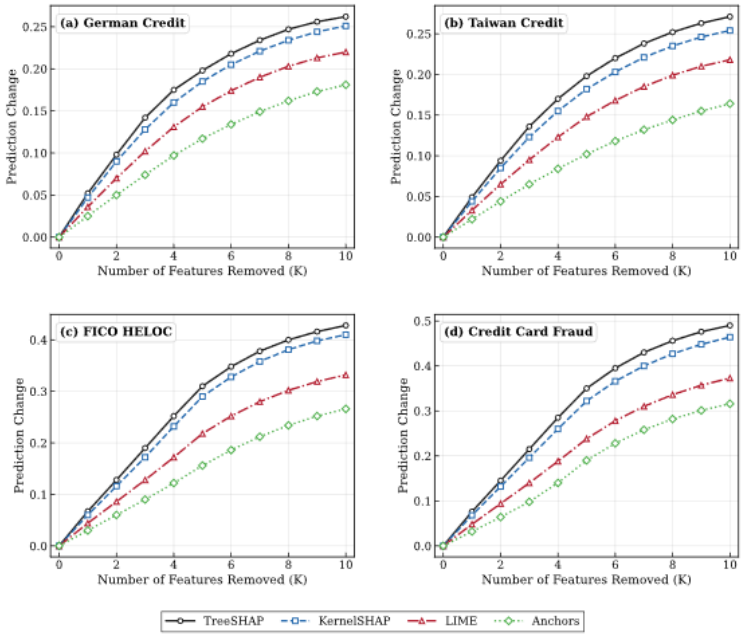

Figure 1. Prediction Change Under Successive Feature Removal (MoRF Order)

Prediction change curves when features are sequentially removed in Most-Relevant-First order on (a) German Credit, (b) Taiwan Credit, (c) FICO HELOC, and (d) Credit Card Fraud using XGBoost. TreeSHAP produces the steepest initial decline across all datasets, with prediction change reaching 0.31 after removing five features on FICO HELOC. KernelSHAP curves closely track TreeSHAP, diverging by less than 0.02 at each removal step. LIME curves show moderate slope with visible variance bands. Anchors curves exhibit the shallowest decline, confirming their lower faithfulness scores in Table 3. The separation between methods is most pronounced on Credit Card Fraud, where TreeSHAP achieves 0.35 prediction change at K = 5 compared to 0.19 for Anchors.

When the Multilayer Perceptron replaces XGBoost as the underlying classifier, Integrated Gradients achieves PGI scores of 0.124, 0.119, and 0.134 on German Credit, Taiwan Credit, and FICO HELOC respectively---competitive with KernelSHAP (0.118, 0.112, 0.128) on the same architecture. This positions Integrated Gradients as a strong candidate for financial applications built on differentiable classifiers, consistent with the axiomatic advantages of gradient path integration identified by Li et al. in their cross-modality benchmarking study [17].

4.2. Stability Comparison

Table 4 presents stability, sparsity, and computational efficiency results averaged across all four datasets on XGBoost. The stability ranking diverges notably from the faithfulness ranking, confirming that these represent genuinely independent quality dimensions.

Table 4. Stability, Sparsity, and Computational Efficiency on XGBoost (Averaged Across Datasets)

Method	Max-Sensitivity (↓)	Gini Coefficient (↑)	Time: German (s)	Time: Taiwan (s)	Time: CCF (s)
TreeSHAP	0.027	0.651	0.003	0.008	0.041
KernelSHAP	0.040	0.626	1.247	2.834	8.521
LIME	0.130	0.593	0.892	1.673	5.147
Anchors	0.099	0.898	3.418	7.256	21.340

Max-sensitivity computed with ε = 0.01 and 50 neighbors. Gini coefficient ranges from 0 (uniform) to 1 (maximally sparse). Time is wall-clock seconds per explanation on a single CPU core.

TreeSHAP achieves the lowest max-sensitivity (0.027), reflecting its deterministic computation that eliminates sampling variance entirely. KernelSHAP attains moderate stability (0.040) owing to the variance-reducing properties of the Shapley estimation procedure. LIME demonstrates the highest max-sensitivity (0.130), nearly five times that of TreeSHAP. This instability arises from the compounding effects of random perturbation sampling and the exponential kernel bandwidth---two sources of variance that do not cancel in expectation. Anchors occupy an intermediate stability position (0.099), where the discrete rule structure dampens small continuous perturbations but produces abrupt changes at rule boundaries (As shown in Figure 2).

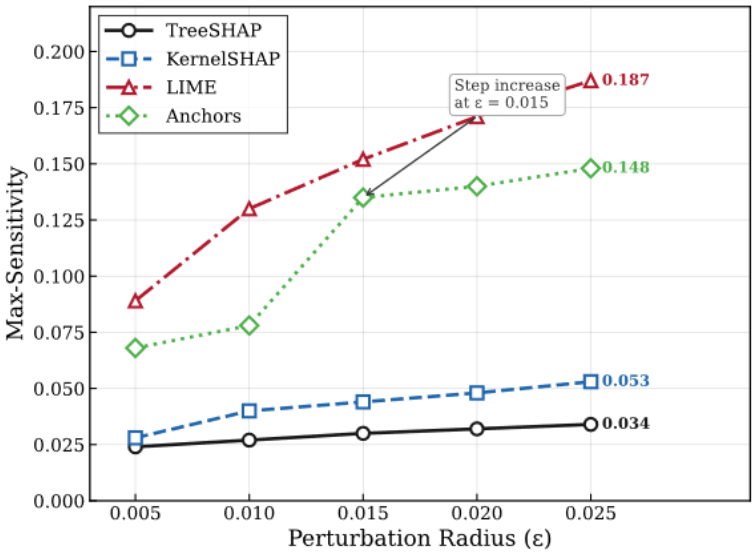

Figure 2. Explanation Stability Across Perturbation Radii

Max-sensitivity as a function of perturbation radius $\varepsilon \in \{0.005, 0.010, 0.015, 0.020, 0.025\}$ on FICO HELOC using XGBoost. TreeSHAP maintains near-constant max-sensitivity below 0.035 across all radii, confirming its deterministic stability. KernelSHAP exhibits a moderate linear increase from 0.028 at $\varepsilon = 0.005$ to 0.053 at $\varepsilon = 0.025$. LIME displays the steepest growth, rising from 0.089 to 0.187 as the perturbation radius increases. Anchors shows a stepped pattern with abrupt sensitivity increases at $\varepsilon = 0.015$, consistent with the discrete nature of rule boundaries.

The stability gap between LIME and Shapley-based methods carries particular significance in financial regulation. When a credit applicant receives a denial, regulatory frameworks often require that the explanation remain consistent if the applicant's profile changes marginally. A max-sensitivity of 0.130 means that LIME explanations can shift substantially under minimal input variation, potentially undermining the legal defensibility of automated decisions.

4.3. Efficiency and Practical Trade-Offs

4.3.1. Computational Cost Analysis

TreeSHAP is the fastest method on all datasets, requiring only 0.003 seconds per explanation on German Credit and scaling to 0.041 seconds on Credit Card Fraud---a three-order-of-magnitude advantage over KernelSHAP (8.521 seconds on CCF). TreeSHAP cost grows sublinearly with dataset size because it depends only on tree structure and feature count. KernelSHAP and LIME scale linearly with feature dimensionality, with KernelSHAP incurring higher per-sample overhead due to the coalition sampling procedure. Anchors is the most expensive method at 21.340 seconds per explanation on Credit Card Fraud, where the 30-dimensional feature space expands the beam search space considerably (As shown in Figure 3).

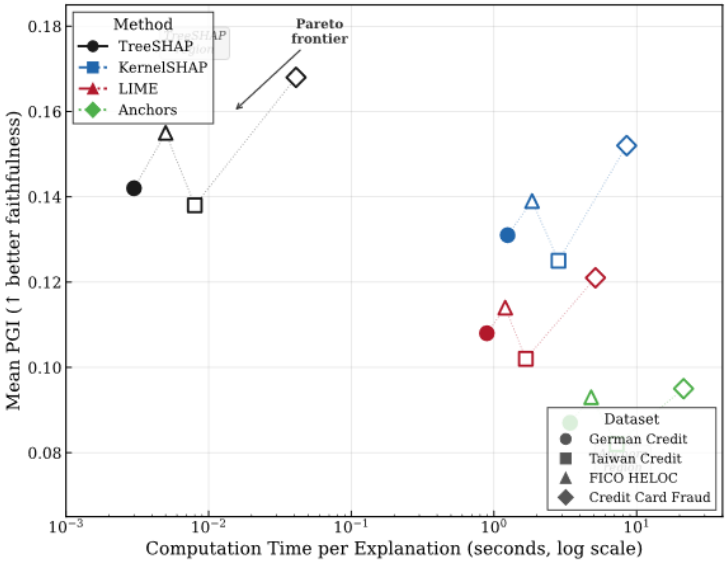

Figure 3. Faithfulness-Efficiency Trade-off Across Methods and Datasets

Scatter plot of mean PGI (y-axis) against log-scaled computation time (x-axis) for each method-dataset combination. TreeSHAP occupies the upper-left region across all four datasets, combining the highest faithfulness (PGI 0.138--0.168) with the lowest computation time (0.003--0.041 s). KernelSHAP clusters in the upper-right quadrant with high faithfulness but substantially greater cost. LIME positions in the center with moderate faithfulness and moderate cost. Anchors occupies the lower-right region with the lowest faithfulness and highest cost. The Pareto frontier is dominated entirely by TreeSHAP points, with KernelSHAP forming a secondary frontier for scenarios where model-agnostic attribution is required.

4.3.2. Multi-Criteria Ranking

Aggregating across faithfulness, stability, and efficiency, TreeSHAP ranks first on tree-based classifiers by a substantial margin. Its only limitation is its restriction to tree architectures, making it inapplicable to deep learning approaches that are gaining traction in financial NLP and sequential modeling tasks. KernelSHAP provides the strongest model-agnostic alternative, achieving the second-highest faithfulness and stability at moderate computational cost. LIME's computational advantage over KernelSHAP diminishes on larger datasets, and its instability penalty (0.130 versus 0.040 max-sensitivity) makes it the least suitable for regulatory contexts among the continuous attribution methods. Anchors occupy a specialized niche: their rule-based format achieves the highest sparsity (Gini = 0.898) and offers unmatched human readability, yet their low faithfulness scores indicate that the rules capture only a coarse approximation of the classifier's decision boundary. Doshi-Velez and Kim argued that practical interpretability requires evaluation along application-grounded, human-grounded, and functionally-grounded dimensions; the metrics in this study address the functionally-grounded dimension, while the sparsity advantage of Anchors may prove decisive when human-grounded evaluations are prioritized [18].

5. Discussion

5.1. Implications for Financial Applications

The empirical evidence assembled in this study points toward a clear practical recommendation for financial institutions deploying tree-based classifiers: TreeSHAP provides the most faithful, stable, and efficient explanations among the four methods evaluated. Its deterministic computation eliminates the reproducibility concerns that plague sampling-based alternatives---a property of direct relevance to model validation teams and external auditors who require identical explanations from identical inputs. The moderate faithfulness gap between KernelSHAP and LIME (0.023--0.031 PGI points) may

appear small in absolute terms, yet the stability difference (0.040 versus 0.130 in max-sensitivity) represents a qualitatively distinct reliability profile that could determine whether an explanation satisfies regulatory scrutiny.

Institutions employing deep tabular architectures such as TabNet or FT-Transformer cannot leverage TreeSHAP and must choose among model-agnostic or gradient-based alternatives. The MLP results in this study indicate that Integrated Gradients achieves faithfulness comparable to KernelSHAP on neural architectures (PGI of 0.124 versus 0.118 on German Credit), with substantially lower computational cost due to direct gradient access. This positions Integrated Gradients as the preferred method when the underlying classifier is differentiable, provided that a meaningful baseline input can be specified---a requirement that warrants careful domain-specific consideration in financial contexts where zero-valued features may not represent a neutral reference point.

The Anchors method, despite ranking lowest on continuous faithfulness metrics, should not be dismissed in practice. Financial compliance officers and non-technical stakeholders frequently prefer discrete rules over continuous attribution vectors because rules map directly to actionable criteria. A hybrid approach combining SHAP values for technical model validation with Anchors rules for customer-facing explanations may satisfy both the accuracy requirements of internal risk management and the accessibility requirements of consumer protection regulations.

5.2. Limitations and Future Directions

Several limitations constrain the generalizability of these findings. The evaluation considers only single-instance local explanations; global explanation quality and the consistency between local and global attributions remain unexamined. The Gaussian perturbation scheme used for infidelity estimation assumes isotropic noise, which may not adequately capture the structured dependencies present in financial features---a limitation that conditional perturbation approaches or copula-based sampling could address in future work. The computational cost measurements reflect single-core execution; TreeSHAP's advantage may narrow under GPU acceleration of KernelSHAP sampling or distributed LIME computation.

Future work should extend this evaluation along several dimensions. Temporal financial datasets where feature distributions shift over time would test whether explanation stability degrades under concept drift, a scenario common in credit scoring where macroeconomic conditions evolve continuously. Incorporating causal constraints into the perturbation protocol could yield faithfulness metrics more aligned with economic reasoning than purely statistical measures, enabling distinction between spurious correlation and genuine causal influence in feature attributions. Connecting specific evaluation metrics to the transparency requirements articulated in Articles 13 and 86 of the EU AI Act would transform abstract quality scores into actionable compliance evidence, bridging the gap between the machine learning and legal communities that both have a stake in trustworthy financial AI.

References

1. Bücker, M., Szepannek, G., Gosiewska, A., and Biecek, P., "Transparency, auditability, and explainability of machine learning models in credit scoring," *Journal of the Operational Research Society*, vol. 73, no. 1, pp. 70--90, 2022.
2. Lundberg, S. M., and Lee, S.-I., "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pp. 4765--4774, 2017.
3. Ribeiro, M. T., Singh, S., and Guestrin, C., "'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 (KDD 2016)*, pp. 1135--1144, 2016.
4. Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., and Lakkaraju, H., "The disagreement problem in explainable machine learning: A practitioner's perspective," *Transactions on Machine Learning Research*, 2024.
5. Alvarez-Melis, D., and Jaakkola, T. S., "Towards robust interpretability with self-explaining neural networks," in *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, pp. 7775--7784, 2018.
6. Sundararajan, M., Taly, A., and Yan, Q., "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, PMLR 70, pp. 3319--3328, 2017.

7. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B., "Sanity checks for saliency maps," in *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, pp. 9505--9515, 2018.
8. Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D. I., and Ravikumar, P. K., "On the (in)fidelity and sensitivity of explanations," in *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pp. 10965--10976, 2019.
9. Hooker, S., Erhan, D., Kindermans, P.-J., and Kim, B., "A benchmark for interpretability methods in deep neural networks," in *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pp. 9734--9745, 2019.
10. Agarwal, C., Krishna, S., Saxena, E., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M., and Lakkaraju, H., "OpenXAI: Towards a transparent evaluation of model explanations," in *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, Datasets and Benchmarks Track, 2022.
11. Han, T., Srinivas, S., and Lakkaraju, H., "Which explanation should I choose? A function approximation perspective to characterizing post hoc explanations," in *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022.
12. Gramegna, A., and Giudici, P., "SHAP and LIME: An evaluation of discriminative power in credit risk," *Frontiers in Artificial Intelligence*, vol. 4, Article 752558, 2021.
13. Bracke, P., Datta, A., Jung, C., and Sen, S., "Machine learning explainability in finance: An application to default risk analysis," *Bank of England Staff Working Paper No. 816*, 2019.
14. Misheva, B. H., Osterrieder, J., Hirsa, A., Kulkarni, O., and Lin, S., "Explainable AI in credit risk management," *arXiv:2103.00949*, 2021.
15. Rong, Y., Leemann, T., Borisov, V., Kasneci, G., and Kasneci, E., "A consistent and efficient evaluation strategy for attribution methods," in *Proceedings of the 39th International Conference on Machine Learning (ICML 2022)*, PMLR 162, 2022.
16. Hedström, A., Weber, L., Krakowczyk, D., Bareeva, D., Motzkus, F., Samek, W., Lapuschkin, S., and Höhne, M. M.-C., "Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations and beyond," *Journal of Machine Learning Research*, vol. 24, no. 34, pp. 1--11, 2023.
17. Li, X., Du, M., Chen, J., Chai, Y., Xiong, H., and Lakkaraju, H., "M4: A unified XAI benchmark for faithfulness evaluation of feature attribution methods across metrics, modalities and models," in *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, Datasets and Benchmarks Track, 2023.
18. Doshi-Velez, F., and Kim, B., "Towards a rigorous science of interpretable machine learning," *arXiv:1702.08608*, 2017.