Journal of Innovative Technologies
https://journalinnotech.com/
Vol 2(2) 2026

Faithfulness, Stability, Agreement, and Complexity of Post-hoc Explanations for Card Fraud Alerts: A Comparative Evaluation of SHAP, LIME, Permutation Importance, and Counterfactuals

Yunfei Gao[1,*] , Chenhui Hao[2]

[1] Electrical Engineering, Northwestern University, IL, USA

[2] Civil Engineering, University of California, CA, USA

* Corresponding Email: gaoyunfei@gmail.com

Abstract

Fraud investigators increasingly rely on post-hoc explanations to justify why a transaction was flagged, yet the reliability of these explanations on actual alert queues is rarely measured. This paper compares three alert-level explanation approaches—SHAP, LIME, and counterfactual explanations—and uses permutation feature importance (PFI) as a retrospective global-importance reference for alerts raised by logistic regression, random forest, and XGBoost classifiers on two public card-transaction datasets: a simulated dataset with documented fraud mechanisms and a real PCA-anonymized dataset. For 200 alerts per classifier, the evaluation covers faithfulness, stability, inter-method agreement, and complexity. Exact SHAP attained the highest PGI-based faithfulness in four of the six settings and showed consistently strong within-setting faithfulness correlations, while maintaining low PGU values.A global PFI ranking matched SHAP for logistic regression but lost alert-level evidence for XGBoost, where its mechanism alignment fell to 0.442. LIME was stable under input perturbations, but its top-three overlap across seeds dropped to 0.453 on the PCA data, and counterfactual importance was the least stable, with run-to-run Jaccard between 0.212 and 0.459. Agreement between explainers was partial and classifier-dependent. The results support reporting explainer configurations and seed stability whenever explanations document fraud alerts.

***Keywords*:** post-hoc explanation; fraud detection; explanation faithfulness; explanation stability

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/)

1. Introduction

1.1. Explanations as Evidence in Fraud Alert Review

Card fraud detection in practice is an alert-review process. A classifier scores every transaction, investigators examine only a small daily budget of the highest-scoring cases, and the labels returned by this review arrive with delay and cover a small fraction of the stream [1]. When an alert is escalated, the institution must be able to state why the activity looked suspicious; in the United States, that justification enters the narrative of a Suspicious Activity Report filed with the Financial Crimes Enforcement Network. Post-hoc attribution tools are attractive in this workflow because they convert a score into a short list of reasons that an investigator can verify and cite.

The same convenience is a source of risk. Explanations of black-box models cannot be perfectly faithful to the computation they summarize, which underlies the argument that high-stakes decisions should rest on inherently interpretable models [2]. Evidence from deployment-like settings is also sobering: an application-grounded study of e-commerce fraud detection found no evidence that the tested post-hoc explanations were useful to analysts [3]. An explanation that does not track the classifier, or that shifts under trivial perturbations, can lend unwarranted authority to a disposition decision.

1.2. Research Gap and Research Questions

A. Gaps in Existing Evaluations

Evaluation practice in explainable AI remains uneven: a systematic review of more than 300 method papers found that one in three relied exclusively on anecdotal evidence [4]. Two further observations matter for alert review. Popular explanation methods frequently disagree about the same prediction, and practitioners often resolve such disagreements with informal heuristics [5]. Perturbation-based explainers can also be inconsistent and unstable, which motivated Bayesian versions of LIME and KernelSHAP that report credible intervals [6]. What is still missing is a side-by-side measurement of how reliable these explainers are on the alerts that fraud classifiers actually raise, under the extreme class imbalance and feature dependence of transaction data.

B. Research Questions

This study asks four questions about alert-level explanations. The first concerns faithfulness: do the features an explainer ranks highest carry more of the classifier's decision than the features it ranks lowest? The second concerns stability: do reference-threshold-preserving perturbations of a transaction, and a change of random seed, leave the leading evidence unchanged? The third concerns agreement: to what extent do the alert-level explanations produced by SHAP, LIME, and counterfactual methods agree with one another and with a retrospective global PFI reference, and does this agreement depend on the classifier and on feature dependence? The fourth concerns complexity: how concentrated is each explanation, and does concentration come at the expense of faithfulness?

1.3. Contributions and Paper Organization

The paper makes three contributions. It defines a reproducible protocol that explains 200 alerts per classifier on a simulated card-transaction dataset with documented fraud mechanisms and on a real, PCA-anonymized card dataset. It reports faithfulness, stability, agreement, and complexity for four explainers across logistic regression, random forest, and XGBoost, including a check of whether explanations recover the injected fraud mechanism. It translates these measurements into guidance for alert review. Section 2 reviews related work, Section 3 describes the experimental setup, Section 4 reports the results, Section 5 discusses implications and limitations, and Section 6 acknowledges data and software providers.

2. Related Work

2.1. Post-hoc Explanation Methods for Tabular Classifiers

A. Additive Attribution with SHAP and LIME

SHAP defines additive feature attributions and shows that Shapley values are the only attributions in this class that satisfy local accuracy, missingness, and consistency, a result that unifies six earlier methods [7]. Its model-agnostic estimator simulates feature absence with a background distribution and treats features as independent, so both the reference data and the dependence structure shape the output. LIME explains a prediction by fitting a sparse linear surrogate to model outputs on perturbed samples weighted by their proximity to the instance [8]. Both methods return signed per-feature scores that read naturally as reasons for an alert, yet they answer different questions: SHAP apportions the gap between a prediction and a reference expectation, whereas LIME approximates local behavior inside a sampling neighborhood shaped by discretization and kernel width.

B. Permutation Importance and Counterfactual Explanations

Permutation importance originates in random forests, where the importance of a variable is the increase in out-of-bag error after its values are randomly permuted [9]. PFI is global: it describes how much a classifier relies on a feature across a dataset rather than why a particular transaction was flagged. A removal-based analysis shows that SHAP, LIME, and permutation tests differ mainly in how features are removed, which model behavior is explained, and how influence is summarized [10]. Counterfactual explanations instead describe a change that would reverse the decision. For tree ensembles, actionable feature tweaking uses the structure of the trees to find such changes [11], while counterfactual generators for mixed-type tabular data can become biased toward particular feature types when features are interdependent [12].

2.2. Evaluating Explanation Quality

Quantitative criteria for feature attributions have been proposed along several complementary axes. Infidelity measures the mismatch between attributions and the output change under input perturbations, and max-sensitivity measures how far an explanation moves when the input moves

slightly [13]. Domain structure adds a further requirement. In financial data, features often form natural groups that conform to regulation, and feature-by-feature Shapley values or integrated gradients may then give inconsistent explanations, whereas group versions of the Shapley value remain consistent [14]. Most studies report such criteria for one explainer at a time on benchmark data, and comparisons that place several explainers on the same alerts while relating faithfulness to stability and agreement remain rare in fraud detection.

2.3. Explainability in Financial Fraud Detection

Explanations in finance are consumed by people who must justify decisions to others. In financial statement audits, SHAP explanations of autoencoder-based anomaly detectors are expressed over encoded variables and are hard for auditors to interpret, which motivated attribute-level aggregation [15]. Fraud models also raise accountability questions beyond accuracy: a bias audit of transaction fraud models showed that certain fairness metrics expose significant bias only after normalization for class imbalance [16]. These studies establish that the quality of an explanation in financial applications depends on the data regime and on its reader. Reports that justify a disposition decision need explanations whose reliability has been checked rather than explanations that merely look plausible. The present work complements these studies with a functional evaluation, without human subjects, of how reliable standard explainers are on fraud alerts.

3. Experimental Setup

3.1. Datasets and Alert Queues

Two public card-transaction datasets with contrasting feature structure are used. The first, HB, is the simulated dataset of the ULB fraud detection handbook, regenerated with the published simulator and parameters and verified against the reported totals of 1,754,155 transactions and 14,681 frauds [17]. Its 15 inputs are the transaction amount, two calendar flags, and 12 customer spending and terminal risk aggregates over 1-, 7-, and 30-day windows, and every fraud records the scenario that produced it: an amount above 220, a compromised terminal, or a compromised customer whose amounts are multiplied by five. The second, ULB, contains 284,807 European card transactions from two days in September 2013, including 492 frauds, with PCA-transformed inputs [18]. HB is split chronologically with a one-week label-delay gap, and ULB is split by time into 70% training and 30% test data. For HB, the evaluation alert set was constructed by selecting the 100 highest-scoring transactions on each test day, while for ULB it consisted of the top 0.25% of test transactions. For subsequent perturbation and counterfactual analyses, HB used a common reference threshold defined as the minimum score among the selected daily alerts, whereas ULB used the cutoff associated with its top-0.25% alert set.Table 1 and Figure 1 summarize both datasets.

Table 1. *Datasets, chronological splits, and alert queues*

Property	HB (simulated)	ULB (real, PCA)
Transactions, train / test	584,546 / 267,930	199,365 / 85,442
Frauds, train / test	5,124 / 2,393	384 / 108
Test fraud rate (%)	0.893	0.126
Input features	15	29
Test window	8 Aug to 4 Sep 2018	last 30% by time
Alerts per classifier	2,800	214
Explained alerts (stability subset)	200 (60)	200 (60)
Pairs with absolute Spearman correlation above 0.5 (%)	10.48	0.49

Source: computed from the regenerated HB data and the OpenML copy of ULB.

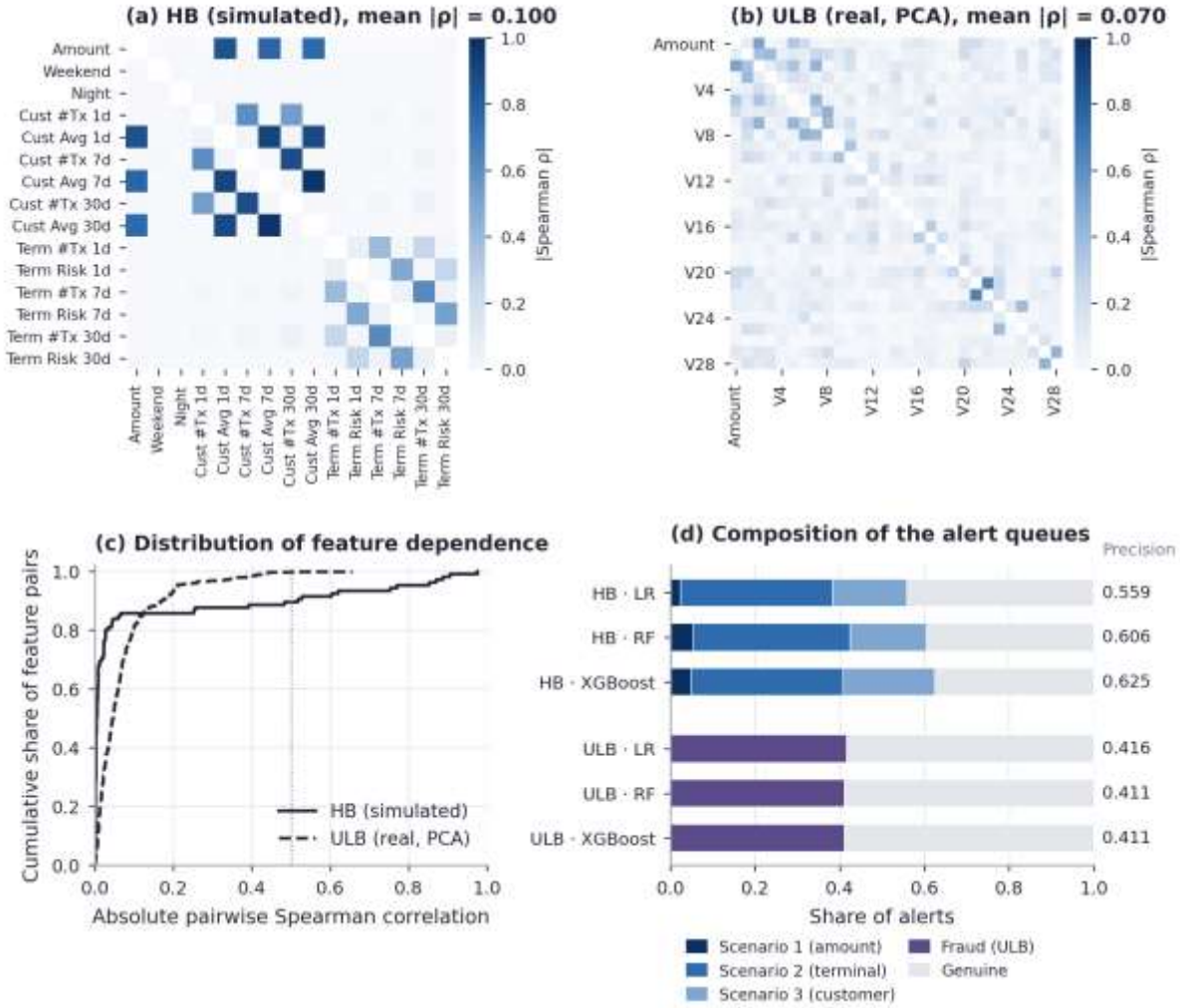


Figure 1. Feature Dependence and Alert Queue Composition in the Two Datasets

(a) Absolute Spearman correlations among the 15 HB features, where the transaction amount, customer average amounts, and window counts form tightly correlated blocks with a maximum of 0.976. (b) The same statistic for the 29 ULB inputs, whose maximum is 0.655. (c) Cumulative distribution of pairwise correlations: 10.48% of HB feature pairs exceed 0.5, against 0.49% of ULB pairs. (d) Composition of each alert queue by generating scenario on HB and by fraud label on ULB; scenario 2 supplies 1,007 of the 2,800 XGBoost alerts, and alert precision is 0.559 to 0.625 on HB and 0.411 to 0.416 on ULB.

3.2. Classifiers and Explanation Methods

A. Fraud Classifiers Under Explanation

Three classifiers common in tabular fraud scoring are explained. Logistic regression (LR) with an L2 penalty and balanced class weights serves as the transparent reference, while a random forest (RF) and gradient-boosted trees trained with XGBoost [19] represent the tree ensembles that dominate practice. Hyperparameters are fixed at the conventional values in Table 2 rather than tuned per explainer, so every explainer faces the same decision function. Detection quality is measured by the area under the precision-recall curve (AUPRC), ROC-AUC, and precision and recall inside the alert queue. Scores are fraud probabilities. For ULB, the lowest score admitted to the top-0.25% queue defines the queue threshold; for HB, the minimum score across all transactions selected by the daily top-100 rule is used as a common reference threshold for the explanation analyses.

B. Explanation Method Configurations

SHAP values are computed exactly, with linear SHAP for LR and interventional TreeSHAP for the ensembles [20], each against 100 background transactions. LIME uses quartile discretization and 5,000 perturbation samples. PFI measures the drop in average precision when a feature is permuted, averaged over ten permutations, and its global ranking is applied to every alert. Counterfactuals follow the random-sampling strategy of DiCE [21] with the settings in Table 2. The share of counterfactuals that change a feature defines its counterfactual importance. Every explainer is run with three seeds, which resample the SHAP background, the LIME perturbations, the PFI permutations, and the counterfactual search.

Table 2. *Classifier and explainer configurations*

Component	Configuration
LR	L2 penalty, C = 1.0, balanced class weights, standardized inputs
RF	200 trees, maximum depth 10, minimum leaf size 20, square-root feature sampling
XGBoost	300 rounds, depth 6, learning rate 0.1, row and column subsampling 0.8

SHAP	Linear SHAP (LR, log-odds); interventional TreeSHAP (RF probability, XGBoost log-odds); 100 background rows
LIME	Tabular LIME with quartile discretization and 5,000 samples
PFI	Decrease in average precision over 10 permutations; 50,000 stratified HB test rows, full ULB test set
Counterfactual	DiCE random search, 10 counterfactuals per alert, target below the dataset-specific reference threshold used in the explanation analysis, calendar flags fixed
Repetition	3 seeds per explainer; 60 alerts with 5 decision-preserving neighbors each

Source: settings of the released experiment code.

3.3. Evaluation Metrics

A. Faithfulness and Complexity

Faithfulness is measured with the prediction-gap metrics of OpenXAI [22]. For an alert and its k highest-ranked features, PGI@k is the mean absolute change in fraud probability when those features are replaced by values drawn jointly from a random training transaction, averaged over 30 draws, and PGU@k applies the same replacement to the k lowest-ranked features; a faithful explanation combines a high PGI with a low PGU. As an auxiliary diagnostic, faithfulness correlation (FC) [23] is computed across 30 random three-feature subsets as the Pearson correlation between the summed feature scores of a subset and the prediction change caused by replacing it. Because SHAP and LIME provide signed local attributions whereas PFI and counterfactual importance are non-directional importance measures, FC is interpreted primarily as an attribution-association diagnostic rather than as a directly comparable faithfulness measure across all four methods. Cross-method faithfulness comparisons therefore rely principally on the prediction-gap metrics PGI and PGU.

is the Pearson correlation, across 30 random three-feature subsets, between the summed attributions of a subset and the probability change caused by replacing it. On HB, mechanism alignment (Hit@3) records whether the top three features of a true-positive alert include a feature that encodes its generating scenario: the amount for scenario 1, terminal risk for scenario 2, and the amount or a customer average amount for scenario 3. Complexity is the entropy of normalized absolute attributions divided by the logarithm of the number of features, together with the number of features needed to reach 80% of the attribution mass. Counterfactuals are further characterized by desiderata surveyed in [24]: validity, sparsity, and plausibility, the last measured as the distance to the nearest genuine training transaction relative to that of genuine test transactions.

B. Stability and Agreement

Input stability is assessed on 60 alerts per classifier. Each alert receives five neighbors built by adding Gaussian noise with a standard deviation of 0.05 feature standard deviations, keeping calendar flags fixed and retaining only neighbors that remain above the common reference threshold used for the explanation analysis.Neighborhood Jaccard@3 compares the top-three feature sets of an alert and its neighbors, and the local Lipschitz estimate divides the largest change in L1-normalized attributions by the corresponding change in standardized inputs, following the stability notion of self-explaining neural networks [25]. Run-to-run Jaccard@3 compares top-three sets across the three seeds. Agreement between two explainers on the same alerts is measured by feature agreement FA@3, the overlap of their top-three sets, and by the Spearman correlation of absolute importances; random rankings give FA@3 values of 0.200 on HB and 0.103 on ULB. Differences in PGI@3 between explainers are tested with paired Wilcoxon signed-rank tests over alerts.

4. Results and Analysis

4.1. Detection Performance and Faithfulness

A. Detection Performance of the Explained Classifiers

All explained classifiers are competent (Table 3). On HB, XGBoost reaches an AUPRC of 0.737 against 0.569 for LR, consistent with evidence that tree-based models remain strong on tabular data [26], while ULB values lie between 0.768 and 0.809. Alert precision between 0.411 and 0.625 means every queue mixes frauds with false alarms.

Table 3. *Detection performance on the test periods*

Dataset	Classifier	AUPRC	ROC-AUC	Alert precision	Alert recall
HB	LR	0.569	0.894	0.559	0.654
HB	RF	0.719	0.900	0.606	0.709
HB	XGBoost	0.737	0.899	0.625	0.732
ULB	LR	0.768	0.979	0.416	0.824
ULB	RF	0.786	0.979	0.411	0.815
ULB	XGBoost	0.809	0.980	0.411	0.815

Source: this study; alert metrics use the queues of Section 3.1.

B. Faithfulness of Attributions and Counterfactuals

SHAP gives strong ranking-based faithfulness in most settings (Table 4, Figure 2). Its PGI@3 is the highest in four of six settings and its PGU@3 never exceeds 0.025. SHAP also shows consistently strong FC values, although FC is treated as an auxiliary diagnostic because its

interpretation is not equivalent for signed attribution methods and non-directional importance measures. Its top three contained a scenario-encoding feature for every explained true-positive HB alert, against a Hit@3 of 0.978 for LIME and 0.914 for counterfactuals with LR. SHAP does not lead everywhere: LIME reaches 0.521 against 0.455 on ULB with LR ($p = 0.002$), and SHAP, PFI, and counterfactual rankings have nearly identical means on ULB with XGBoost (0.380, 0.381, and 0.376).

Table 4. *Faithfulness of alert explanations: PGI@3 and faithfulness correlation (FC)*

Dataset	Method	LR PGI@3	RF PGI@3	XGBoost PGI@3	LR FC	RF FC	XGBoost FC
HB	SHAP	0.654	0.488	0.560	0.893	0.930	0.839
HB	LIME	0.436	0.453	0.502	0.653	0.558	0.606
HB	PFI	0.608	0.424	0.266	0.463	0.350	0.103
HB	Counterfactual	0.393	0.474	0.510	0.258	0.703	0.413
ULB	SHAP	0.455	0.270	0.380	0.743	0.958	0.766
ULB	LIME	0.521	0.151	0.352	0.517	0.431	0.464
ULB	PFI	0.366	0.222	0.381	0.263	0.614	0.452
ULB	Counterfactual	0.110	0.168	0.376	0.097	0.641	0.461

Source: this study; means over 200 explained alerts per classifier.

A global ranking is a weak substitute for alert-level evidence. PFI matches SHAP with LR on HB (0.608, $p = 0.147$) but falls to 0.266 with XGBoost, because its global top three omit terminal risk and its Hit@3 is 0.000 for all 63 scenario-2 alerts, echoing evidence that variable importance varies across well-performing models [27]. Counterfactual importance is least faithful on ULB with LR (PGI@3 of 0.110, faithfulness correlation of 0.097) yet competitive on HB with RF (0.474). Counterfactual validity was 1.000 in every setting except RF on ULB (0.807), and counterfactuals changed 1.62 to 4.93 features on average. Despite this sparsity, ULB counterfactuals lay 9.11 to 14.41 times the typical nearest-neighbor distance of genuine transactions, against 3.52 to 4.87 for the alerts, consistent with unjustified counterfactuals [28]. Optimization-based search for tree ensembles finds closer counterfactuals [29], so these random-search results are conservative.

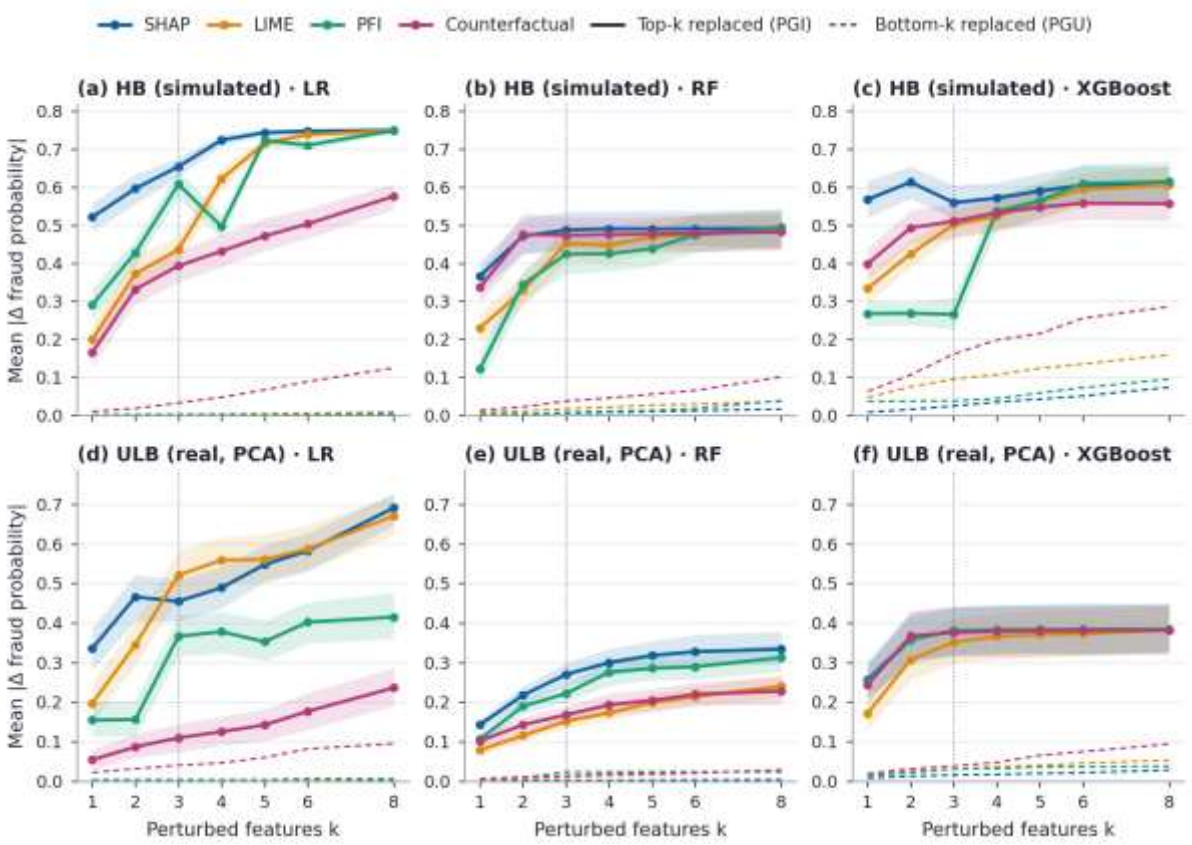


Figure 2. Prediction-Gap Faithfulness Curves Across Classifiers and Datasets

Mean absolute change in fraud probability when the k highest-ranked (PGI) or lowest-ranked (PGU) features are replaced, with 95% bootstrap bands, for (a) LR, (b) RF, and (c) XGBoost on HB and (d) LR, (e) RF, and (f) XGBoost on ULB. SHAP leads at k = 3 in (a), (b), (c), and (e), with PGI@3 of 0.654, 0.488, 0.560, and 0.270. In (c), PFI stays near 0.266 until a fourth feature is replaced, and counterfactual rankings give the highest PGU@3 in five panels, up to 0.162.

4.2. Stability of Explanations

SHAP and LIME are stable under reference-threshold-preserving perturbations, whereas counterfactual importance is not (Figure 3). Neighborhood Jaccard@3 spans 0.915 to 0.990 for SHAP and 0.890 to 0.988 for LIME but only 0.370 to 0.876 for counterfactuals, and SHAP has the lowest median local Lipschitz estimate in five of six settings. Across seeds, PFI is invariant, SHAP stays between 0.919 and 0.972, and counterfactual importance reaches only 0.212 to 0.459, while LIME falls to 0.453 with RF and 0.566 with XGBoost on ULB. With the seed fixed, the same RF explanations reach a neighborhood Jaccard@3 of 0.988, locating the instability in perturbation sampling, which S-LIME addresses through statistically controlled sampling [30].

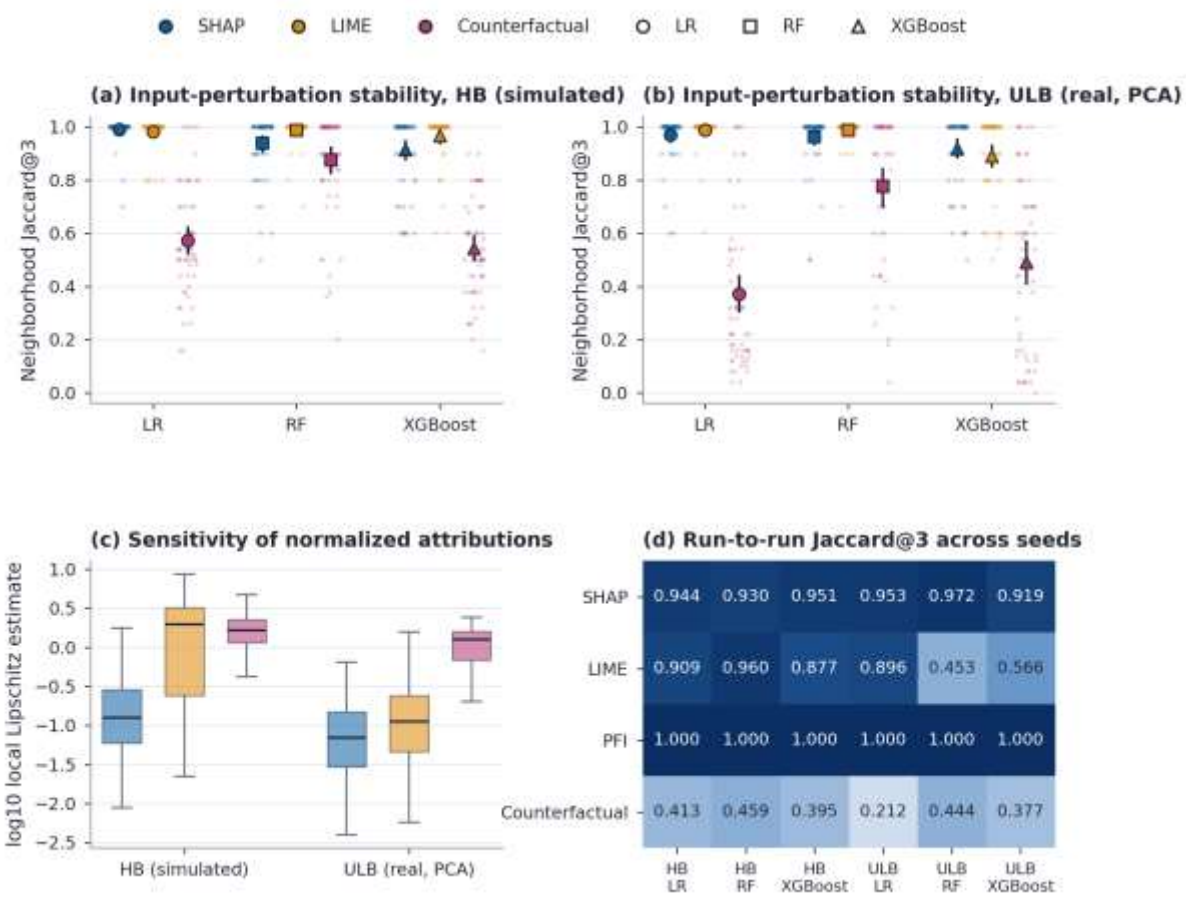


Figure 3. Input-Perturbation and Run-to-Run Stability of Alert Explanations

(a) Neighborhood Jaccard@3 of HB alerts, with per-alert values, means, and 95% bootstrap intervals; counterfactual means range from 0.543 with XGBoost to 0.876 with RF. (b) The same measure on ULB, where counterfactual importance drops to 0.370 with LR. (c) Local Lipschitz estimates pooled over classifiers, with the lowest median for SHAP on both datasets. (d) Run-to-run Jaccard@3 across three seeds: 1.000 for PFI, 0.919 to 0.972 for SHAP, down to 0.453 for LIME, and 0.212 to 0.459 for counterfactuals.

4.3. Agreement and Complexity

A. Inter-method Agreement

Agreement between explainers is partial and classifier-dependent (Figure 4). SHAP and LIME share 0.485 to 0.720 of their top three features on HB and 0.470 to 0.597 on ULB, and the Spearman correlation of their absolute importances spans 0.295 to 0.817, as the function-approximation view predicts for methods approximating a classifier over different neighborhoods and losses [31]. Counterfactual importance is the outlier: with LR on ULB it shares 0.113 of the SHAP and 0.082 of the LIME top three, near chance. Mean pairwise FA@3 spans 0.469 to 0.538 on HB and 0.296 to 0.520 on ULB, so the stronger dependence of HB did not lower agreement, although differing random baselines limit this comparison. Dependence still matters, because

Shapley estimates that assume independence evaluate unrealistic feature combinations when features are correlated [32].

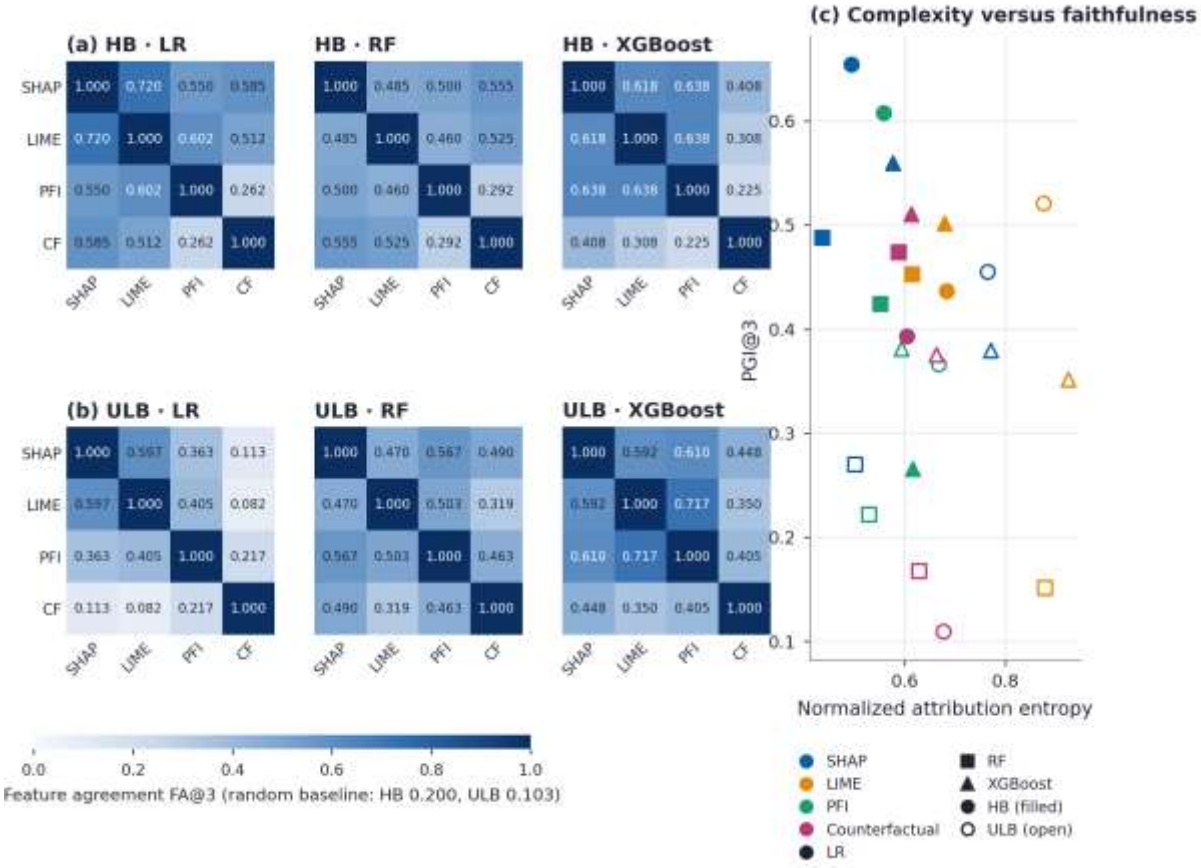


Figure 4. Inter-Method Agreement and the Relation Between Complexity and Faithfulness

(a) Mean FA@3 among SHAP, LIME, PFI, and counterfactual (CF) rankings on HB, ranging off the diagonal from 0.225 (PFI and CF, XGBoost) to 0.720 (SHAP and LIME, LR). (b) The same matrices for ULB, where CF agreement with SHAP and LIME under LR is 0.113 and 0.082 against a random baseline of 0.103. (c) Normalized attribution entropy against PGI@3 for all 24 combinations; LIME reaches the highest entropy (0.925), while SHAP attains both the lowest HB entropy (0.436) and the highest PGI@3 (0.654).

B. Explanation Complexity

LIME is the most diffuse explainer in every setting, with normalized entropy of 0.615 to 0.925 and up to 17.52 of 29 ULB features needed for 80% of the attribution mass. SHAP is the most concentrated on all HB classifiers, needing 2.62 to 3.45 features, and also the most faithful there, so concentration did not cost faithfulness; diffuse LIME rankings accompanied a weak prediction gap on ULB with RF (0.151).

5. Discussion and Future Work

5.1. Implications for Alert Review and Documentation

The results point to a division of labor among explainers rather than a single winner. For the tree ensembles, exact interventional SHAP combined the highest faithfulness correlation in every setting with high stability under both input perturbations and seed changes, and its top three contained a scenario-encoding feature for every explained true-positive alert of the simulated data. This makes it the most defensible default for recording why an alert fired. LIME explanations should not be quoted from a single run: their leading features were stable under small input changes but varied across seeds for the tree ensembles on the higher-dimensional data, so repeated runs with a reported overlap statistic are a minimal safeguard. A global PFI ranking is a poor proxy for alert-level reasons. It matched SHAP for logistic regression but missed the terminal-risk evidence for all 63 scenario-2 alerts explained for XGBoost, a silent failure that would pass unnoticed in a disposition narrative. Counterfactuals answer a different and useful question, namely what would have kept a transaction below the alert threshold, yet their importance rankings were the least stable, and on PCA features the counterfactuals lay far from genuine transactions. They suit probing a decision boundary during an investigation better than serving as the stated basis of a report. A practical documentation standard could record, for each escalated alert, the explainer and its configuration, the leading features with their attributions, and the overlap of those features across repeated runs. Model risk teams could apply the same four measurements as acceptance tests before an explainer enters an investigation workflow and, where confirmed fraud typologies exist, verify that the leading features of true frauds reflect the known mechanism. Such checks cost little compared with investigation time that an unreliable explanation can misdirect.

5.2. Limitations and Future Work

The evidence comes from one simulated and one PCA-anonymized dataset, so neither reproduces the mix of readable and masked features found in production fraud systems, and the simulated fraud mechanisms are simpler than real fraud. Faithfulness was measured by marginal replacement, which can create unrealistic transactions when features are dependent, and mechanism alignment used a lenient top-three criterion with few scenario-1 alerts. Classifier hyperparameters and alert budgets were fixed, counterfactuals came from a random search, and input stability was assessed on 60 alerts per classifier. Only three classifier families were examined, so the relative standing of the explainers could change for neural scoring models, whose attributions rely on different approximations. Explanation reliability may differ in queues where false positives dominate the reviewed cases, and the ULB test period covers only the final part of a two-day window, which limits claims about behavior over time. The study also involved no investigators, so it measures whether explanations are reliable rather than whether they improve decisions. Future work should repeat the protocol on large real transaction data with interpretable features, compare conditional Shapley estimators and optimization-based counterfactual search, test group-level attributions for correlated feature blocks, and run application-grounded studies in

which investigators use explanations to draft suspicious activity narratives. Extending the protocol to anti-money-laundering alerts and graph-based detectors would show whether this division of labor holds beyond card fraud.

References

[1] Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2018). Credit card fraud detection: A realistic modeling and a novel learning strategy. *IEEE Transactions on Neural Networks and Learning Systems, 29*(8), 3784–3797. https://doi.org/10.1109/TNNLS.2017.2736643

[2] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[3] Amarasinghe, K., Rodolfa, K. T., Jesus, S., Chen, V., Balayan, V., Saleiro, P., Bizarro, P., Talwalkar, A., & Ghani, R. (2024). On the importance of application-grounded experimental design for evaluating explainable ML methods. *Proceedings of the AAAI Conference on Artificial Intelligence, 38*(19), 20921–20929. https://doi.org/10.1609/aaai.v38i19.30082

[4] Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., & Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Computing Surveys, 55*(13s), 1–42. https://doi.org/10.1145/3583558

[5] Krishna, S., Han, T., Gu, A., Wu, S., Jabbari, S., & Lakkaraju, H. (2024). The disagreement problem in explainable machine learning: A practitioner's perspective. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=jESY2WTZCe

[6] Slack, D., Hilgard, S., Singh, S., & Lakkaraju, H. (2021). Reliable post hoc explanations: Modeling uncertainty in explainability. *Advances in Neural Information Processing Systems, 34*. https://papers.nips.cc/paper_files/paper/2021/hash/4e246a381baf2ce038b3b0f82c7d6fb4-Abstract.html

[7] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*. https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[8] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

[9] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[10] Covert, I., Lundberg, S., & Lee, S.-I. (2021). Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research, 22*(209), 1–90. https://jmlr.org/papers/v22/20-1316.html

[11] Tolomei, G., Silvestri, F., Haines, A., & Lalmas, M. (2017). Interpretable predictions of tree-based ensembles via actionable feature tweaking. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 465–474). Association for Computing Machinery. https://doi.org/10.1145/3097983.3098039

[12] Panagiotou, E., Heurich, M., Landgraf, T., & Ntoutsi, E. (2024). TABCF: Counterfactual explanations for tabular data using a transformer-based VAE. In *Proceedings of the 5th ACM International*

Conference on AI in Finance (pp. 274–282). Association for Computing Machinery. https://doi.org/10.1145/3677052.3698673

[13] Yeh, C.-K., Hsieh, C.-Y., Suggala, A. S., Inouye, D. I., & Ravikumar, P. (2019). On the (in)fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems, 32*. https://papers.nips.cc/paper_files/paper/2019/hash/a7471fdc77b3435276507cc8f2dc2569-Abstract.html

[14] Chen, D., Chen, J., & Ye, W. (2024). Why groups matter: Necessity of group structures in attributions. In *Proceedings of the 5th ACM International Conference on AI in Finance* (pp. 63–71). Association for Computing Machinery. https://doi.org/10.1145/3677052.3698667

[15] Müller, R., Schreyer, M., Sattarov, T., & Borth, D. (2022). RESHAPE: Explaining accounting anomalies in financial statement audits by enhancing SHapley Additive exPlanations. In *Proceedings of the Third ACM International Conference on AI in Finance* (pp. 174–182). Association for Computing Machinery. https://doi.org/10.1145/3533271.3561667

[16] Kamalaruban, P., Pi, Y., Burrell, S., Drage, E., Skalski, P., Wong, J., & Sutton, D. (2024). Evaluating fairness in transaction fraud models: Fairness metrics, bias audits, and challenges. In *Proceedings of the 5th ACM International Conference on AI in Finance* (pp. 555–563). Association for Computing Machinery. https://doi.org/10.1145/3677052.3698666

[17] Le Borgne, Y.-A., Siblini, W., Lebichot, B., & Bontempi, G. (2022). *Reproducible machine learning for credit card fraud detection: Practical handbook*. Université Libre de Bruxelles. https://github.com/Fraud-Detection-Handbook/fraud-detection-handbook

[18] Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. In *2015 IEEE Symposium Series on Computational Intelligence* (pp. 159–166). IEEE. https://doi.org/10.1109/SSCI.2015.33

[19] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

[20] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

[21] Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 607–617). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372850

[22] Agarwal, C., Krishna, S., Saxena, E., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M., & Lakkaraju, H. (2022). OpenXAI: Towards a transparent evaluation of post hoc model explanations. *Advances in Neural Information Processing Systems, 35*. https://papers.nips.cc/paper_files/paper/2022/hash/65398a0eba88c9b4a1c38ae405b125ef-Abstract-Datasets_and_Benchmarks.html

[23] Bhatt, U., Weller, A., & Moura, J. M. F. (2020). Evaluating and aggregating feature-based model explanations. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* (pp. 3016–3022). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2020/417

[24] Guidotti, R. (2024). Counterfactual explanations and how to find them: Literature review and benchmarking. *Data Mining and Knowledge Discovery, 38*(5), 2770–2824. https://doi.org/10.1007/s10618-022-00831-6

[25] Alvarez-Melis, D., & Jaakkola, T. S. (2018). Towards robust interpretability with self-explaining neural networks. *Advances in Neural Information Processing Systems, 31*.

https://papers.nips.cc/paper_files/paper/2018/hash/3e9f0fc9b2f89e043bc6233994dfcf76-Abstract.html

[26] Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems, 35*. https://papers.nips.cc/paper_files/paper/2022/hash/0378c7692da36807bdec87ab043cdadc-Abstract-Datasets_and_Benchmarks.html

[27] Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research, 20*(177), 1–81. https://jmlr.org/papers/v20/18-760.html

[28] Laugel, T., Lesot, M.-J., Marsala, C., Renard, X., & Detyniecki, M. (2019). The dangers of post-hoc interpretability: Unjustified counterfactual explanations. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence* (pp. 2801–2807). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2019/388

[29] Lucic, A., Oosterhuis, H., Haned, H., & de Rijke, M. (2022). FOCUS: Flexible optimizable counterfactual explanations for tree ensembles. *Proceedings of the AAAI Conference on Artificial Intelligence, 36*(5), 5313–5322. https://doi.org/10.1609/aaai.v36i5.20468

[30] Zhou, Z., Hooker, G., & Wang, F. (2021). S-LIME: Stabilized-LIME for model explanation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (pp. 2429–2438). Association for Computing Machinery. https://doi.org/10.1145/3447548.3467274

[31] Han, T., Srinivas, S., & Lakkaraju, H. (2022). Which explanation should I choose? A function approximation perspective to characterizing post hoc explanations. *Advances in Neural Information Processing Systems, 35*. https://papers.nips.cc/paper_files/paper/2022/hash/22b111819c74453837899689166c4cf9-Abstract-Conference.html

[32] Aas, K., Jullum, M., & Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence, 298*, Article 103502. https://doi.org/10.1016/j.artint.2021.103502