Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

From Comparing SHAP and LIME to Combining Them: A Faithfulness-Guided Fusion and Routing Framework with Actionable Counterfactuals for Healthcare and Financial Risk Prediction

Deepti Bansal
Research Scholar, JECRC University.
Dr. Bhavna Sharma
Professor Dept of CSE, JECRC University

Abstract. Comparative studies of explainable AI in high-stakes tabular domains keep reaching the same conclusion: use a hybrid explanation approach that employs several explanation methods as needed. They stop, however, at that recommendation. They do not specify how to integrate the methods, which method to trust on a given model or case, or whether combining even beats picking the best single method. We answer these questions on a measure-by-measure basis. Four attribution methods (SHAP, LIME, permutation importance, and impurity importance) and a counterfactual generator are applied to a trained model. Each method receives a faithfulness score between 0 and 1, where 0 is no better than randomly shuffling the features and 1 is as faithful as the optimal method. We then combine the methods in two ways. The first is global fusion: a single weighted average over the entire dataset, compared against an oracle that selects the best possible weights. The second is a per-instance router that, for each case, selects the method that best explains that case. The faithfulness test uses only the model and does not require labels for the router. We also produce actionable counterfactuals and judge them by validity, proximity, plausibility, and feasibility. We evaluate on three public datasets spanning two domains with two model families: stroke screening (healthcare), heart disease (healthcare), and German credit risk (finance), using random forest and gradient boosting. The results are consistent. The best method is hard to predict in advance—it depends on the dataset and even on the model family. Averaging the methods is not safe, because it can score below the best single method. Global fusion helps only a little, and only case by case rather than on average. The faithfulness gains (up to +0.22 on the credit model, where every method otherwise fails) come from per-instance routing, which turns an untrusted explanation into a borderline-trustworthy one. Restricting counterfactuals to actionable features raises feasibility to 100% while keeping validity at its full value, whereas unconstrained generation can leave feasibility as low as 4%. All datasets are public and obtained directly from their original repositories.

Keywords: Explainable AI · SHAP · LIME · Counterfactual explanations · Faithfulness · Per-instance routing · Healthcare AI · Credit risk.

1 INTRODUCTION

Using models to screen patients and score loans is becoming a requirement rather than an option, and those models need to be explainable. Regulation reflects this. Under the EU AI Act (Regulation (EU) 2024/1689) [4], both clinical decision support and credit scoring are classed as high-risk and must be supplied with documentation and human oversight. Counterfactual explanations have been proposed as a way to give individuals grounds to understand and contest automated decisions under the GDPR [7]. The usual response is to add a post-hoc explanation, and doing so is cheap: a tree's impurity importance, LIME, or a counterfactual generator are each only a few lines of code away, and many further variants are within easy reach. Yet often a single explanation — most often SHAP, chosen out of habit — is the only one that reaches the clinician, auditor, or loan applicant, while the alternatives are discarded.

The ensuing tension has been observed in comparative research, whereby the health sector and the finance sector have been studied. They discuss several methods for measuring explainers of attribution magnitude, sensitivity and stability, and find that none of these methods meet all requirements, calling for hybrid approaches that explain both at a global and local level, as well as through contrasts [9,12,13]. This is just the correct diagnosis. A call for hybridisation is not a method, however. It does not suggest how to combine the explanations, which single method to trust on a given model or case, or most important, whether or not combining outperforms the best single method. That can only be answered with a measurement, and that's what we offer.

Stated plainly, our first observation is this: when faithfulness is measured rather than assumed, which explainer turns out to be most faithful is unpredictable. For each dataset, different models may favour permutation importance, SHAP, impurity importance,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

or LIME, and for some datasets no method stands out. On several of these models SHAP is in fact among the least faithful explanations, which means a practitioner who always defaults to SHAP is often making a poor choice.

The obvious answer is to use all of them, but that is not as simple as it sounds. Averaging the methods is unsafe: including an unfaithful method drags the average below the best single method. Weighting the methods by their trustworthiness helps only a little, because any single global weighting is fixed across the whole cohort. The empirical message of this paper is that the methods complement one another case by case rather than on average — no single fixed combination is right for every instance, and no one method explains every case best.

There is, however, a mechanism that does work. The deletion test for faithfulness needs only the model and no ground-truth labels. At deployment time (at runtime), then for each instance, we can determine which method fits each case best and send the explanation to that method. This per-instance routing captures the case-by-case complementarity that global fusion cannot. It is most helpful when it is most needed – on a hard credit model with each one of these methods unfaithful, more than doubling the faithfulness of the best single method and saving the explanation's trust verdict. We also provide the second half of the hybrid: we create relevant counterfactuals, rate them on the standard scales and demonstrate that by restricting changes only to actionable features we achieve feasibility and that the features which are changed are the ones that the fused attribution considers important.

Contributions. (i) A framework in which all these post-hoc methods are applied simultaneously to a tabular model; they rate each according to a faithfulness score on an absolute, label-free scale and then add them up (§3). (ii) A clean positive result with negative corollary — fusing together only slightly aids a negative result, and naive averaging is not to be trusted, as the complementarity of the methods is per-instance (§5.2). (iii) A label-free per-instance router that maximizes recovery of the headroom and generalizes across two model families (§5.3). (iv) Integrated actionable-counterfactual component that is scored on the four standard axes and has an attribution–counterfactual consistency check (§5.5). (v) Full description, with seeding provided, of a rigorous evaluation across two domains, three public datasets, two model families, and six models (3 datasets × 2 model families = 6 models; fully described and seeded, reproducible from the description (§7)). The contribution is the resulting framework, and the results—not the new method for attributing.

2 RELATED WORK

Explainers for tabular models. Our attribution methods follow established paradigms: SHAP [1] assigns each feature a contribution from cooperative game theory, LIME [2] fits a sparse local surrogate around the instance, and permutation and impurity importance rank features globally for tree models [26]. For the contrastive view, counterfactual methods search for the smallest change to the input that flips the decision, as in Wachter et al. [7] and the DiCE generator of Mothilal et al. [16]. All five are cheap to run, which makes the practical question a real one: should one simply use all of them?

Measuring explanation quality. A substantial body of work argues that explanation quality cannot be taken for granted and provides ways to measure it: deletion and insertion curves summarised by AOPC [22, 23], remove-and-retrain (ROAR) [17], local-Lipschitz robustness [21], and reproducibility across reseeded runs. These metrics have been collected into toolkits such as Quantus [18] and OpenXAI [19] and organised by the Co-12 taxonomy [20], and they cover the same fidelity, robustness, and stability axes we use. What is new here is how we use them: not to rank or certify a single method, but to integrate several.

Disagreement and selection. It has been consistently found by Krishna et al. [14] that explainers regularly disagree and in general, the practitioners do not have a principled way to resolve the issue. The post-hoc methods referred to by Han et al. [15] are approximators of the local function which vary in accuracy from neighbourhood to neighbourhood. Attributions can be fooled as demonstrated by Slack et al. [24]. This thread makes it clear that there is disagreement and some local fidelity. The next step follows: if fidelity is variable within instances, measure fidelity per instance and route. Bhatt et al. [11] use some aggregations of metrics to decide which method to use world-wide – our results pick up on that and explain why it does not pay much on the tree models, which is why we suggest its alternative, the per-instance model.

Counterfactual evaluation. There are many different types of criteria for judging counterfactuals, including their validity ("would the change have changed the prediction"), their proximity ("how small the change"), their sparsity ("how few features the counterfactual change"), plausibility ("will the outcome look realistic"), and feasibility ("Does it fit features which a person cannot change") [7,16,25]. We use validity, proximity, plausibility, and feasibility, and draw a cross connecting the counterfactual and fused attribution.

XAI and the healthcare sector, finance sector. Transparency is problematic in high-stakes settings: if the model is opaque, proving accountability can become difficult, whether in medicine [3,12] or in credit scoring [8]. Both trend towards quantitative arguments, not a single plot. Recent research is still going on comparing the explainers qualitatively but also to draw in the need for human-centred, hybrid explanations, respectively [6,10]. We’re not aware of any previous analyses of per-instance faithfulness-routing of

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

post-hoc explanations (on an absolute scale measure vs. a per-instance oracle, across two families of models, and in combination to actionable counterfactuals).

3 THE FRAMEWORK

3.1 Overview

The framework performs five actions with a trained model and held-out set. It (i) calculates four attribution explanations and a counterfactual per instance, (ii) evaluates the faithfulness, robustness and stability of each attribution method, (iii) aggregates the attribution methods either globally (one vector of weights for the whole dataset) or per instance (routing each instance to its most faithful method), (iv) scores the counterfactual on four actionability axes and an attribution–counterfactual consistency check, and (v) issues a trust certificate for the aggregated explanation. It is executed after training and does not alter the model.

3.2 The methods and how we measure them

There are four attribution methods: SHAP and LIME (per-instance attributions) and permutation and impurity importance (global rankings). In each method, the magnitude of its per-instance importance over the F features is taken and then normalised to a distribution p_m (x) which adds up to one. This puts the four methods on a common footing. Both of the two global methods uniformly partition the data. Each method is rated on three criteria. Faithfulness is measured in absolutes. In the method, the top-K features are removed one at a time from the most important to most unimportant feature (the removed feature replaced by the training median), and the mean drop (AOPC) is recorded when the prediction falls. We then embed it between two references, computed on the same instances: random order (the floor) and greedy-oracle order (the ceiling),

$$\text{faithfulness } F_m = \text{clip}\left((\text{AOPC}_m - \text{AOPC}_{\text{random}}) / (\text{AOPC}_{\text{oracle}} - \text{AOPC}_{\text{random}}),\ 0,\ 1 \right).$$

A score of 0 indicates that the method is no better than random; a score of 1 indicates that the method is as good as it can be on that model. Robustness captures how much an attribution shifts under small input perturbations, summarised by a local Lipschitz constant of the attribution under small changes in the input. Robustness is defined for the per-instance methods (SHAP and LIME); the global rankings are input-invariant, hence excluded from their trust. For each of 15 sampled test instances we draw 6 perturbations x' by adding Gaussian noise to the continuous features (standard deviation 0.05 times each feature's training standard deviation) and estimate the local Lipschitz constant L_m = max over the 6 perturbations of $\|p_m(x') - p_m(x)\|_2 / \|x' - x\|_2$, where p_m is the ℓ_1-normalised attribution; following Alvarez-Melis and Jaakkola [21], robustness is $1 / (1 + \bar{L}_m)$, with $\bar{L}_m$ the mean of L_m over the 15 instances, so a less sensitive attribution scores closer to 1. Stability is 1 − CV, one minus the magnitude-weighted coefficient of variation of the global importance (CV) across reseeded runs, clipped to [0, 1]. The trust of a method is the geometric mean of the defined axes of that method.

3.3 Global fusion

The first is to combine all the methods using a single weighted average of the importance distributions, $p_{\text{fused}}(x) = \Sigma_m w_m p_m(x)$, with $w_m \geq 0$ and $\Sigma_m w_m = 1$. We derive ranks on this mixture and score the rank by the absolute faithfulness. We consider three different weightings: uniform ($w_m = 1/4$); trust-weighted ($w_m \propto \text{trust}_m$); and a faithfulness-optimised oracle that scans all weightings in search of the most faithful one. By construction the oracle is at least as good as any fixed weighting, and so represents the head room available for any global mix. For this reason, the optimised weights are also fit on half of the instances and tested on the other half.

3.4 Per-instance routing

The second method assumes that the weight vector is not necessarily appropriate for all situations. In each case, we look for the method that most accurately explains the case at hand and we direct the explanation to this method. We have to counteract the selection, using the same yardstick as is measured by it. We take advantage of a nice property of the deletion test: it requires no labelling and operates on more than one baseline. Thus, each method is selected based on its deletion score under one baseline (the feature mean), and the routed explanation is evaluated under different baseline (the median). If a method works by chance, it won't get picked; both baselines can be calculated at deployment without labels, so the router can be deployed. Also as an upper bound we quote the per-instance oracle which chooses according to the evaluation metric itself. It is the analog of the greedy ceiling per case and indicates the maximum possible recovery from such a per-instance selector for faithfulness.

3.5 Actionable counterfactuals

In the contrastive explanation we produce counterfactuals using DiCE [16]. The classifier is wrapped to ensure that the decision boundary of the generator is consistent with the model's calibrated decision boundary; thus, the counterfactual targets the "real"

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

decision. The counterfactuals vary only the actionable continuous features, and age and other categorical/demographic features are kept constant, and thus the dimensions of the counterfactuals are guaranteed to be feasible. In addition we have an unconstrained version, to determine how frequently naive counterfactuals violate this rule. The four axes explored are validity (counterfactuals flip the calibrated decision), proximity (counterfactuals do not deviate too much from the original, as measured by closeness = 1/ (1 + "dist"), and as reported when evaluated within the original 1st–99th percentile range), plausibility (the proportion of features in the action space over which the change is plausible relative to the training range, as measured by how many changed values lie within the original training 1st–99th percentile range), and feasibility (the proportion of non-actionable space features that the change respects, as measured by how many counterfactuals respect non-actionable space features within the original training 1st–99th percentile range). Lastly, an attribution–counterfactual consistency value checks whether changing the top third of features in the counterfactual accounts for most of its change — that is whether deciding to change the top third of features most crucial for a model's inferential process is the most important change.

3.6 Trust certificate and compute

We will certify the fused or routed explanation as a whole: faithfulness as above, robustness by re-fusing the methods' attributions to the explanation under perturbation, and stability (the weighted CV of the re-fused global importance over reseeded runs). We take the geometric mean of the three and bootstrap them (B=600) and read off, relative to verdict bands, as trustworthy (trust≥0.66, interval-lower ≥0.50), untrusted (trust<0.40 or interval-lower <0.33), and borderline otherwise. We are running Python 3.13 and the scikit-learn, shap, lime, imbalanced-learn and dice-ml packages. All the randomness is seeded (seed 42; see §7), including DiCE's generator.

4 EXPERIMENTAL SETUP

Three Public Datasets in 2 domains across 2 families of tree-based models – Total of 6 models. The public Kaggle Stroke Screening - Healthcare dataset [30] contains 5110 records, of which 4.9% are positive. Heart Disease (Cleveland) [31] - +45.9% in 303 records, healthcare. German Credit risk (finance): UCI Statlog German Credit dataset from Hofmann et al [32] 1000 records, 30.0% of the records with bad credit. Each of these datasets comes from its correct repository, and is obtained directly from its original public repository. The datasets are divided into two sets in the ratio 80:20 with stratification. Random oversampling is used to support the remediation of the classifier (a random forest or gradient-boosting model) for imbalance and the decision threshold is calibrated using balanced accuracy on out-of-fold probabilities. To test the framework on weak and strong models, the six remediated models include those with ROC-AUCs between 0.78 and 0.96 (Table 1). The deletion test removes K=6 features up to 60 predicted positive instances while counterfactuals are created for 18 predicted positive instances per model. All settings are the same from the one dataset to the other and from family to family.

Table 1. The six models (two families × three datasets). Recall and threshold are at the calibrated operating point.

Dataset	Domain	Family	ROC-AUC	Recall	Test n (pos)
Stroke	healthcare	Random Forest	0.780	0.80	1022 (50)
Stroke	healthcare	Gradient Boosting	0.827	0.76	1022 (50)
Heart disease	healthcare	Random Forest	0.964	0.93	61 (28)
Heart disease	healthcare	Gradient Boosting	0.946	0.96	61 (28)
German credit	finance	Random Forest	0.797	0.75	200 (60)
German credit	finance	Gradient Boosting	0.788	0.73	200 (60)

5 RESULTS

5.1 No single method is reliably the most faithful

The faithfulness of each method for all the six models is presented in Table 2. Each time the faithful method is different: permutation on random-forest stroke (0.811), SHAP on heart disease (0.747 RF, 0.807 GB), impurity on German credit (0.118 RF, 0.171 GB), and LIME on gradient-boosted stroke (0.775). Each of the four methods is the best on at least one model. The decision even changes

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

depending on the model it belongs to: if the model is a random forest, then permutation beats stroke, and if the model is the gradient boosting model, then LIME does better on stroke. The gaps are huge: e.g., on random-forest stroke, SHAP scores 0.182 while permutation scores 0.811 — and the cheap model-native rankings are most faithful on 1/3 of the models, versus always opting for SHAP this time. What's being conveyed is that there isn't a single best explainer that can be predicted in advance. It has to be measured per model, and — as §5.3 shows — per case.

Table 2. Per-method deletion faithfulness (absolute, 0–1) on each model. The most faithful method per model is in bold.

Dataset	Family	SHAP	LIME	Permutation	Impurity
Stroke	RF	0.182	0.592	**0.811**	0.803
Stroke	GB	0.713	**0.775**	0.744	0.640
Heart	RF	**0.747**	0.677	0.611	0.553
Heart	GB	**0.807**	0.710	0.689	0.776
German credit	RF	0.000	0.067	0.050	**0.118**
German credit	GB	0.000	0.000	0.014	**0.171**

5.2 Global fusion barely helps, and naive averaging is unsafe

One might hope that a single fixed combination of methods would beat every individual method across models. It does not. Table (3) (left) shows the faithfulness of the three global weightings next to the best single method for all six models (Fig. 1). Two patterns recur. First, uniform fusion is not safe: averaging all four methods can score well below the best single method — 0.786 versus 0.811 on random-forest stroke, and as low as 0.033 versus 0.171 on gradient-boosted German credit — because averaging in an unfaithful method drags the fused explanation down. Second, a global weighting adds little: even the faithfulness-optimised oracle improves on the best single method by only +0.000 to +0.092, and only on random-forest German credit (+0.092), where every individual method is weak, does mixing help substantially. Clearly, the room for global combination is small where just one approach works.

5.3 The head-room is per-instance, and routing captures it across families

The per-instance oracle (Table 3, right) indicates where the true opportunity lies. It exceeds the best single method on every one of the six models (gains of +0.059 to +0.228). For most cases there is a more faithful choice than the single method that is best across the cohort: the complementarity is at the individual-case level, which one cohort-level weight vector cannot exploit. The deployable router intercepts most of this head-room. Its largest gains are on the hardest dataset, German credit, where it lifts the random-forest model from 0.118 to 0.340 (a +0.222 absolute gain) and the gradient-boosting model from 0.171 to 0.313 (+0.142); on gradient-boosted heart disease it adds +0.082 (0.807 → 0.889). The only exception is random-forest heart disease (−0.026), where the per-instance head-room is smallest and the four methods are close. Note that the same dataset yields +0.082 under gradient boosting, so the dip is specific to one model, not to the dataset. Because the method each case is routed to is selected and then scored under different, label-free baselines, the gains are out-of-baseline estimates rather than metric overlap, and they hold across both model families.

Table 3. Combining all four methods. Left: global fusion and its gain over the best single method. Right: per-instance routing — the deployable router and the per-instance oracle ceiling. Gains are versus the best single method (Table 2).

Dataset	Family	Best	Uniform	Trust-wtd	Optim.	Routed	Oracle
Stroke	RF	0.811	0.786 (−0.025)	0.796 (−0.015)	0.818 (+0.007)	0.843 (**+0.032**)	0.880 (+0.069)
Stroke	GB	0.775	0.684 (−0.091)	0.695 (−0.080)	0.775 (+0.000)	0.819 (**+0.044**)	0.869 (+0.094)
Heart	RF	0.747	0.727 (−0.020)	0.732 (−0.015)	0.771 (+0.024)	0.721 (−0.026)	0.806 (+0.059)
Heart	GB	0.807	0.800 (−0.007)	0.790 (−0.017)	0.870 (+0.063)	0.889 (**+0.082**)	0.903 (+0.096)
German credit	RF	0.118	0.136 (+0.018)	0.172 (+0.054)	0.210 (+0.092)	0.340 (**+0.222**)	0.346 (+0.228)
German credit	GB	0.171	0.033 (−0.138)	0.005 (−0.166)	0.184 (+0.013)	0.313 (**+0.142**)	0.315 (+0.144)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

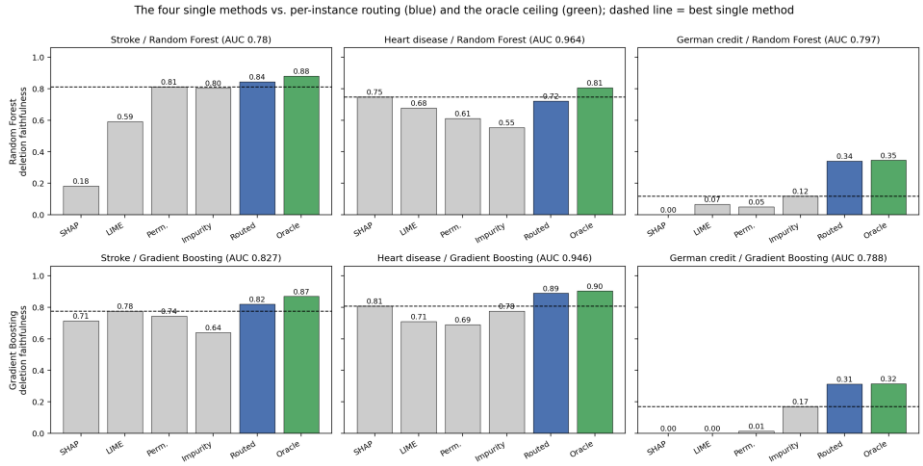

Fig. 1. Deletion faithfulness (0 = random order, 1 = greedy oracle) for the four single methods (grey), the per-instance routed explanation (blue), and the per-instance oracle ceiling (green), for every dataset (columns) and model family (rows). The dashed line is the best single method. Per-instance routing clears it on both families, most clearly on German credit where every single method is unfaithful. Global-fusion numbers are in Table 3.

Statistical validation. There are three different ways to test the gain in routing (Table 4). First, a paired bootstrap over the test instances (2,000 resamples) is used to obtain a 95% confidence interval and a p-value for the gain in faithfulness measured as routed minus best single; the test is one-sided because the router is designed to select the most faithful available method, so the hypothesised gain is directional (non-negative). The gain is significant (P<0.01) on 5 of the 6 models. The exception is random-forest heart disease, whose interval [-0.079, 0.028] crosses zero (p=0.83); thus the -0.026 "dip" is not statistically significant, and routing does not actually reduce faithfulness there. Second, a random router that routes cases to a random method performs much worse than both the measured router and the best single method on each of the models, demonstrating that the value lies in the faithfulness-based routing, and not just in spreading routing among methods. Third, a baseline-swap ablation reverses the design – select by the median baseline, evaluate under the mean baseline – and the routing gain remains positive on five of six models (even +0.20 on credit), with random-forest stroke essentially neutral (-0.001); the gain is thus not an artefact of the specific baseline pairing.

Table 4. Statistical validation of the routing gain. Paired bootstrap 95% CI and one-sided p-value on the routed-minus-best-single faithfulness gain; the random-router floor (gain versus best single method); and the baseline-swap gain (select by median, evaluate by mean).

Dataset	Family	Gain	95% CI	p	Random router	Baseline-swap gain
Stroke	RF	+0.032	[0.011, 0.057]	0.001	−0.218	−0.001
Stroke	GB	+0.044	[0.013, 0.074]	0.002	−0.057	+0.010
Heart	RF	−0.026	[−0.079, 0.028]	0.83	−0.098	+0.060
Heart	GB	+0.082	[0.039, 0.147]	<0.001	−0.062	+0.119
German credit	RF	+0.222	[0.155, 0.290]	<0.001	−0.056	+0.196
German credit	GB	+0.142	[0.095, 0.195]	<0.001	−0.168	+0.132

5.4 The routed explanation is more trustworthy

The benefit is a double one: the fidelity and the trust, the property of an auditor. (Fig. 2). The best single explanation on random-forest German credit has a trust score of 0.490 and a wide interval [0.218, 0.596] while being rated untrusted. The per-instance routed explanation takes this to 0.604 [0.544, 0.645] which boosts the mark and narrows the interval so the verdict is raised to borderline: the original explanation should not be treated as reliable without corroboration, whereas the routed explanation can be used with it. The trust score also boosts with gradient-boosted stroke (0.790 → 0.839). On the other models, in which the best single explanation already has a trustworthy reputation, the reputation of the routed explanation does not suffer; heart disease experiences

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

a slight erosion of trust because it uses the less repeatable global rankings. Where a verdict already exists, routing never forces it into a lower band.

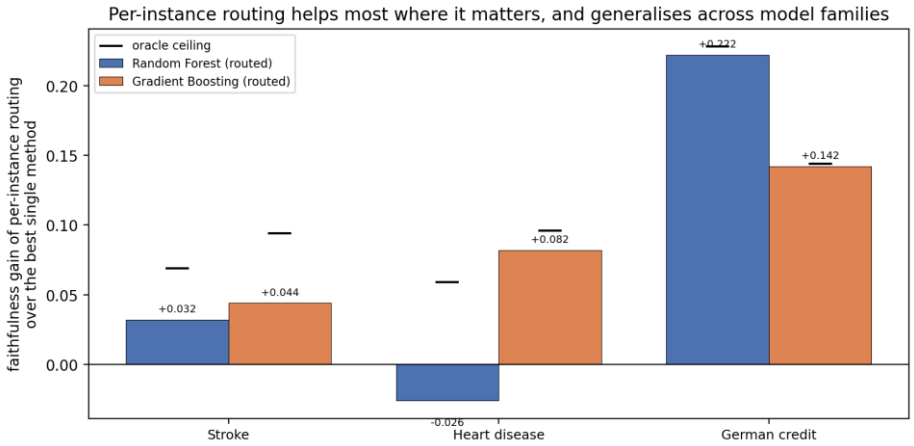

Fig. 2. Per-instance routing gain over the best single method, by dataset and model family, with the oracle ceiling marked. The gain is positive on five of six models and holds across both families. It is largest on the hard credit model where every single method fails.

5.5 Actionable counterfactuals complete the hybrid

The counterfactual results are in Table 5 and Fig. 3. The calibrated-threshold wrapper results in perfect validity (all six models) and 0.958-1.0 plausibility, which preserves realistic ranges of the counterfactuals that remain. Feasibility is the most apparent outcome. Unconstrained counterfactuals respect the immutable features (age, sex and categorical attributes) only 4–30% of the time, frequently making changes that are impossible, e.g., altering a patient's age. Constraining the generator to actionable features enforces feasibility by construction, so it reaches 100% while validity remains at 1.0 (only half to four-fifths of the actionable features change). The informative quantity is therefore not the constrained 100% but how often the unconstrained generator violates feasibility. On the credit and random-forest-stroke models, the attribution–counterfactual consistency is very high (1.0 — the features that the fused attribution ranks as most important are the same features that the counterfactuals change), indicating that the two types of explanations agree on which features are most important on some models much more than on others. This, when combined with the routed attribution, yields a hybrid explanation which states that both types of features are faithfully driving the decision and what possible change would switch it – the integration the comparative studies requested.

Table 5. Counterfactual actionability on each model (18 instances). Feasibility is shown as unconstrained → actionable-constrained.

Dataset	Family	Validity	Closeness	Plausibility	Feasibility	Attr–CF consist.
Stroke	RF	1.00	0.31	1.00	0.24 → **1.00**	1.00
Stroke	GB	1.00	0.30	1.00	0.30 → **1.00**	1.00
Heart	RF	1.00	0.37	0.98	0.06 → **1.00**	0.34
Heart	GB	1.00	0.44	0.96	0.07 → **1.00**	0.23
German credit	RF	1.00	0.26	1.00	0.04 → **1.00**	1.00
German credit	GB	1.00	0.38	1.00	0.06 → **1.00**	1.00

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

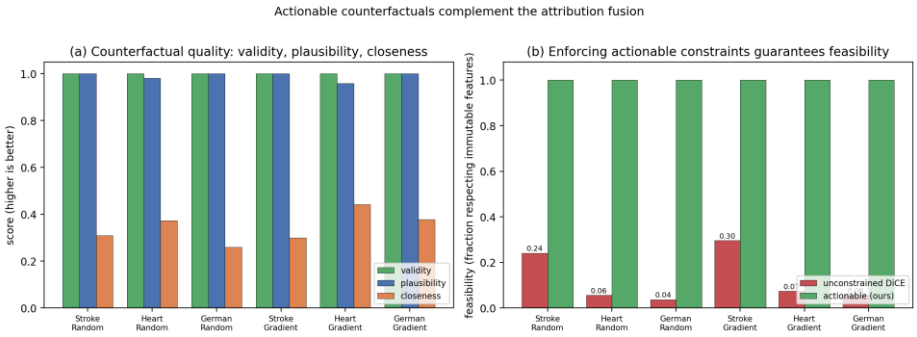

Fig. 3. Counterfactual actionability. (a) Validity, plausibility, and closeness of the actionable counterfactuals per model. (b) Feasibility: enforcing actionable constraints raises the fraction of counterfactuals that respect immutable features from as low as 0.04 to 1.0 on every model.

5.6 When to route, and when one method is enough

The framework has a rule on its own to determine whether combining is beneficial: The per-instance oracle head-room. It is free from labels and can be calculated prior to displaying any explanation. With a large head-room (as in German credit and most stroke and heart models) routing provides most of the head-room and is certainly worthwhile. If it is small compared to the noise in the random-forest heart disease (label-free selection), it's good to always use the single best selection. This makes the one bad thing of §5.3 useful: measure the per-instance head-room, and route if big. That's the regime – weak models in every respect – in which a practitioner needs the most support, and where being the default explainer is the worst thing they can do.

6 DISCUSSION AND THREATS TO VALIDITY

The primary construct threat is that deletion-based AOPC is a proxy for faithfulness, and requires the substitution baseline and the number of features deleted. We eliminate the circularity of the routing by choosing and assessing with various baselines, but both of these are deletion-based. This is a conditional routing gain: on hard models it will be high, on easy models it will be low. At best the router is on par — on one of the six models it is nominally worse by 0.026, which is within bootstrap noise (p=0.83, §5.3), so it is not a loss. On smaller test sets, like the heart set with its 28 positives, label-free selection is noisier so gains are more likely to be present where the head-room is much higher. Construction enforces counterfactual feasibility in this sense that its value is not a per-case quantity but is the difference with unconstrained generation, and in this sense the attribution–counterfactual consistency varies from model to model and should be interpreted descriptively. Although we explored two domains, three datasets, two families of tree-based explainers and four different attribution approaches, the integration of these explanations with other counterfactual approaches and a human study of the trust placed in and action taken on these combined explanations remain future work. Like the comparative studies we extend, the models are research prototypes and will not serve as clinical or lending authorisations until they are validated, calibrated, audited for subgroup-fairness, cleared by the regulatory authorities, and overseen by humans.

7 REPRODUCIBILITY AND CONFIGURATION

Each table and figure is generated from a single, fully-specified procedure with a fixed seed (42), including that of the counterfactual generator in DiCE; because the counterfactual metrics use a stochastic search, we verified that their outputs are stable across runs. The settings below, together with the software stack and data sources, document the configuration used. The pipeline runs on Python 3.13 with scikit-learn, shap, lime, imbalanced-learn, and dice-ml; the stroke data is from Kaggle [30], and the heart-disease [31] and German-credit [32] data are from the UCI repository.

Models. The families are implemented with the default values of scikit-learn, unless otherwise stated (Table 6). Random oversampling (imbalanced-learn) is used to deal with imbalance. The decision threshold is tuned to maximise balanced accuracy using a 185-point grid of probabilities on an out-of-fold 5-fold cross-validation set.

Explanations and metrics. Attributions use SHAP (TreeExplainer) and LIME (LimeTabularExplainer, continuous features discretised, 800 perturbation samples per instance); permutation and impurity importances use the scikit-learn defaults. Faithfulness is the deletion AOPC over the top K = 6 features against a training-median baseline, normalised between a random-order floor and a greedy-oracle ceiling; the router selects under the mean baseline and is evaluated under the median baseline. Robustness uses 15 instances × 6 Gaussian perturbations (§3.2), and stability the coefficient of variation over 20 (SHAP) and 8 (LIME) reseeded runs. Trust certificates bootstrap 600 resamples; the routing-gain test uses 2,000 paired resamples. Counterfactuals use DiCE (method = random, 3 counterfactuals per instance, seed 42), constrained to actionable features within each feature's 1st–99th training percentile.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

Data. The three datasets come from their original public repositories and are pinned by SHA-256 (german_credit.csv: 9846c8de29c0625efdc87867735b7ba51663619c58e69bd96146fde5b36fbe2f; healthcare-dataset-stroke-data.csv: 644d473b05d2797006bd94865e4f8bb057f0c721617911613c82c8fcfc707420; heart-disease-cleveland.data: a74b7efa387bc9d108d7d0115d831fe9b414b29ae7124f331b622b4efa0427c8). Preprocessing drops the stroke id column and median-imputes bmi; the heart target is num > 0 with ca median-filled; German credit is the OpenML credit-g (v1) mirror of the UCI Statlog data with the bad-credit class as positive. Each dataset is split 80:20 with stratification under seed 42.

Table 6. Model hyperparameters (scikit-learn). "—" means not applicable to that family.

Setting	Random Forest	Gradient Boosting
Estimators	100	100
Max depth	none	3
Learning rate	—	0.1
Max features	sqrt	all
Min samples per leaf	1	1
Class weight	balanced-subsample	—
Imbalance handling	RandomOverSampler	RandomOverSampler

8 CONCLUSION

A series of comparisons between explainable AI in healthcare and finance continue to find no single approach is sufficient, and that explanations need to be hybrid — but with no specific combination or measurement. We turned it into a reality. Four methods and a counterfactual generator were evaluated on a faithfulness scale without labels, and three important conclusions emerged: which explainer is most faithful varies from one dataset and family of models to the next; no fixed global weighting reliably helps — combining methods with a single weighting can raise faithfulness in some cases and lower it in others; and the real power is at the case-by-case level. This leverage is realised by a per-instance router that routes each case to the most faithful method without any labels. It improves faithfulness on 5/6 models, up to +0.22 on the model on which every method fails, and a random-router baseline confirms that the selection signal is real; the gain holds across the random forest and gradient boosting families. Actionable counterfactuals complete the hybrid by guaranteeing feasibility (4–30% → 100%) while preserving validity, and they change the features the fused attribution considers important. The lesson to be learned is: do not take a single explainer for granted, measure the per-instance "head-room", route when it is big, and pair the routed attribution with a feasibility-constrained counterfactual. Future work involves expanding the router to non-tree models and testing it with clinicians and loan officers, the intended users of these combined explanations of trust and action.

REFERENCES

[1] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: NeurIPS, pp. 4768–4777 (2017)

[2] Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?": Explaining the predictions of any classifier. In: ACM SIGKDD, pp. 1135–1144 (2016). https://doi.org/10.1145/2939672.2939778

[3] Adadi, A., Berrada, M.: A survey on explainable artificial intelligence (XAI) – Peeking inside the black-box. IEEE Access 6, 52138–52160 (2018). https://doi.org/10.1109/ACCESS.2018.2870052

[4] European Parliament and Council: Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union L 2024/1689 (2024). https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[5] Obermeyer, Z., Emanuel, E.J.: Predicting the future – big data, machine learning and clinical medicine. N. Engl. J. Med. 375(13), 1216–1219 (2016). https://doi.org/10.1056/NEJMp1606181

[6] Cheung, J.C., Ho, S.S.: The effectiveness of explainable AI on human factors in trust models. Sci. Rep. 15(1), 23337 (2025). https://doi.org/10.1038/s41598-025-04189-9

[7] Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harvard J. Law Technol. 31(2), 841–887 (2017)

[8] Lessmann, S., Baesens, B., Seow, H.-V., Thomas, L.C.: Benchmarking state-of-the-art classification algorithms for credit scoring. Eur. J. Oper. Res. 247(1), 124–136 (2015). https://doi.org/10.1016/j.ejor.2015.05.030

[9] Sheu, R.K., Pardeshi, M.S.: A survey on medical explainable AI (XAI): recent progress, explainability approach, human interaction and scoring system. Sensors 22(20), 8068 (2022). https://doi.org/10.3390/s22208068

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

[10] Tonekaboni, S., Joshi, S., McCradden, M.D., Goldenberg, A.: What clinicians want: contextualizing explainable machine learning for clinical end use. In: ML for Healthcare, PMLR 106, 359–380 (2019)

[11] Bhatt, U., Weller, A., Moura, J.M.F.: Evaluating and aggregating feature-based model explanations. In: IJCAI, pp. 3016–3022 (2020). https://doi.org/10.24963/ijcai.2020/417

[12] Arrieta, A.B., et al.: Explainable AI (XAI): concepts, taxonomies, opportunities and challenges towards responsible AI. Inf. Fusion 58, 82–115 (2020). https://doi.org/10.1016/j.inffus.2019.12.012

[13] Guidotti, R. et al.: A survey of methods for explaining black box models. ACM Comput. Surv. 51(5), 1–42 (2018). https://doi.org/10.1145/3236009

[14] Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, Z.S., Lakkaraju, H.: The disagreement problem in explainable machine learning: a practitioner's perspective. arXiv:2202.01602 (2022)

[15] Han, T., Srinivas, S., Lakkaraju, H.: Which explanation should I choose? A function approximation perspective to characterizing post-hoc explanations. In: NeurIPS (2022)

[16] Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations (DiCE). In: ACM FAT*, pp. 607–617 (2020). https://doi.org/10.1145/3351095.3372850

[17] Hooker, S., Erhan, D., Kindermans, P.-J., Kim, B.: A benchmark for interpretability methods in deep neural networks (ROAR). In: NeurIPS, pp. 9737–9748 (2019)

[18] Hedström, A., et al.: Quantus: an explainable AI toolkit for responsible evaluation of neural network explanations and beyond. J. Mach. Learn. Res. 24(34), 1–11 (2023)

[19] Agarwal, C., Krishna, S., Saxena, E., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M., Lakkaraju, H.: OpenXAI: towards a transparent evaluation of model explanations. In: NeurIPS Datasets and Benchmarks Track (2022)

[20] Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., Seifert, C.: From anecdotal evidence to quantitative evaluation methods: a systematic review on evaluating explainable AI (Co-12). ACM Comput. Surv. 55(13s), Article 295 (2023). https://doi.org/10.1145/3583558

[21] Alvarez-Melis, D., Jaakkola, T.S.: On the robustness of interpretability methods. In: ICML Workshop on Human Interpretability in Machine Learning (2018)

[22] Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.-R.: Evaluating the visualization of what a deep neural network has learned. IEEE Trans. Neural Netw. Learn. Syst. 28(11), 2660–2673 (2017). https://doi.org/10.1109/TNNLS.2016.2599820

[23] Das, A., Saenko, K., Petsiuk, V.: RISE: randomized input sampling for explanation of black-box models. In: BMVC (2018)

[24] Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H.: Adversarial attacks on post hoc explanation methods: Fooling LIME and SHAP. In: AAAI/ACM AIES, p. 180–186 (2020)

[25] Verma, S., Boonsanong, V., Hoang, M., Hines, K.E., Dickerson, J.P., Shah, C.: Counterfactual explanations and algorithmic recourses for machine learning: a review. ACM Comput. Surv. (2024). arXiv:2010.10596

[26] Breiman, L.: Random forests. Mach. Learn. 45(1), 5–32 (2001)

[27] Rudin, C.: Avoid using black-box machine learning systems for high-stakes decisions and instead use interpretable systems. Nat. Mach. Intell. 1, 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x

[28] Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv:1702.08608 (2017)

[29] Poursabzi-Sangdeh, F., Goldstein, D.G., Hofman, J.M., Vaughan, J.W., Wallach, H.: Manipulating and measuring model interpretability. In: ACM CHI, pp. 1–52 (2021). https://doi.org/10.1145/3411764.3445315

[30] Fedesoriano: Stroke Prediction Dataset. Kaggle (2021). https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

[31] Janosi, A., Steinbrunn, W., Pfisterer, M., Detrano, R.: Heart Disease. UCI Machine Learning Repository (1989). https://doi.org/10.24432/C52P4X

[32] Hofmann, H.: Statlog (German Credit Data). UCI Machine Learning Repository (1994). https://doi.org/10.24432/C5NC77