

Explainable Artificial Intelligence in Clinical Decision Support: A Comparative Evaluation of SHAP, LIME, Attention and Permutation Attribution for Cardiovascular Risk Prediction

Yazan Roux^{1*}, Akinyi Dlamini²

¹University of Lagos, NIGERIA

²The University of Jordan, JORDAN

Abstract

Background. Explainable artificial intelligence is increasingly presented as the mechanism that will make machine learning acceptable in clinical decision support. Yet the leading attribution methods are rarely compared against one another on the same model and data, so it is unclear whether they agree, whether they are stable, and what they cost. **Objective.** To quantify the agreement, stability, fidelity and computational cost of four attribution methods — SHAP, LIME, learned feature attention and permutation feature importance — applied to cardiovascular risk prediction. **Methods.** Four classifiers (L2-regularised logistic regression, random forest, gradient-boosted trees and a purpose-built attention-gated network) were trained on the UCI Cleveland heart-disease cohort (n = 302 after de-duplication; 54.3% prevalence) under stratified ten-fold cross-validation. Attributions were generated on a held-out 30% partition and compared on four axes: rank agreement (Spearman ρ , Kendall τ), run-to-run stability (25 repeated explanations per case), fidelity (area under a per-instance deletion curve, AUDC) and wall-clock cost per explained patient. **Results.** Discrimination was statistically indistinguishable across model families: random forest 0.909 ± 0.059 , logistic regression 0.899 ± 0.060 , gradient boosting 0.885 ± 0.054 and the attention network 0.876 ± 0.059 AUC. The four attribution methods disagreed substantially, with pairwise Spearman ρ ranging from 0.582 (SHAP vs. permutation) to 0.846 (SHAP vs. LIME); serum cholesterol was ranked third most important by SHAP and last of thirteen by permutation importance. LIME attributions varied across identical repeated runs with a mean coefficient of variation of 0.279 and a top-five Jaccard similarity of 0.799, whereas SHAP and attention were exactly reproducible. SHAP was the most faithful (AUDC 0.662) and LIME the costliest at 13.48 ms per patient, some 6,700 times slower than reading attention weights. **Conclusions.** Explanation method is not an implementation detail: it changes which risk factor a clinician is told mattered. Attention weights were fast but collapsed onto a single feature and were the least faithful of the post-hoc methods; LIME was both the slowest and the least reproducible. Since a regularised logistic regression matched every black box here, the interpretability of the model itself remains the strongest available guarantee.

Keywords: Explainable AI; SHAP; LIME; Attention Mechanisms; Clinical Decision Support; Cardiovascular Risk Prediction; Model Interpretability; Explanation Stability

Introduction

Cardiovascular disease remains the leading cause of death worldwide, and machine learning models for risk stratification now routinely match or exceed traditional risk scores on retrospective cohorts

[1], [2]. Their adoption into clinical workflows has nonetheless been slow, and the reason most often given is opacity. A clinician asked to act on a model's output reasonably wants to know why that output was produced, and regulators increasingly require that the reasoning be articulable [3].

The field's answer has been post-hoc explainability. SHAP, grounded in cooperative game theory, attributes a prediction to its input features by computing each feature's average marginal contribution across coalitions [4]. LIME fits a sparse interpretable surrogate to the model's behaviour in the local neighbourhood of the instance being explained [5]. Attention mechanisms take a different route, learning a weighting over inputs during training and exposing those weights as an explanation. Permutation feature importance measures the degradation in performance when a feature is randomly shuffled. All four are widely deployed in clinical machine learning papers, frequently in combination and frequently without comparison [6], [7], [8].

This is the gap the present study addresses. If two attribution methods applied to the same model and the same patient nominate different risk factors, then the explanation shown to the clinician is partly an artefact of a tooling decision made by the engineer. That concern is not hypothetical: the disagreement between explanation methods has been documented as a general phenomenon in machine learning [9], and adversarial work has shown that both LIME and SHAP can be manipulated to conceal a model's reliance on a sensitive feature [10]. Rudin's argument that high-stakes decisions should use inherently interpretable models rather than explained black boxes rests substantially on this instability [3].

Most clinical XAI studies report explanations as a qualitative addendum — a SHAP summary plot, a sentence observing that the important features are clinically plausible — rather than evaluating the explanation itself. Where quantitative treatment does appear, it is usually confined to a single method [11], [12]. Systematic head-to-head comparison on identical models and data, with stability and fidelity measured rather than assumed, remains uncommon [1-8].

Related Work

Attribution methods

SHAP unifies several earlier attribution schemes under the Shapley value, the unique allocation satisfying local accuracy, missingness and consistency [4]. Its principal practical drawback is cost: exact computation is exponential in the number of features, and although TreeSHAP reduces this to polynomial time for tree ensembles, model-agnostic variants remain expensive. LIME samples perturbed instances around the point of interest, weights them by proximity and fits a sparse linear surrogate [5]. Because the perturbations are drawn randomly, LIME explanations are stochastic — a property acknowledged in the applied literature but rarely quantified [8].

Attention has a contested status as explanation. Jain and Wallace demonstrated that attention weights frequently fail to correlate with gradient-based importance and that adversarially constructed attention distributions can leave predictions unchanged, concluding that attention is not explanation [13]; Wiegrefe and Pinter replied that the claim depends on what one asks of an explanation [14]. Permutation feature importance is model-agnostic and cheap to reason about, but is global rather than local and is known to behave erratically under correlated predictors [10-14].

Explainable AI in cardiovascular decision support

Applied work in this area is abundant. Recent frameworks combine SHAP, LIME and permutation importance to interpret cardiovascular risk models, reporting that the identified risk factors align with clinical knowledge [6]. Others introduce explainability stability, defined as the consistency of feature-importance rankings across models rather than across explanation methods, and use SHAP as the single interpretability tool [11]. A dual-explainability framework combining LIME with permutation importance selected the latter over SHAP explicitly on computational-cost grounds, while noting that LIME is sensitive to random sampling and may vary between runs [8]. Ensemble approaches have begun to report Kendall's τ and overlap@10 as explanation-stability metrics across folds and subgroups [7].

Two patterns are visible across this literature. First, method selection is frequently justified by convenience rather than by measured properties. Second, where multiple methods are applied, agreement between them is asserted qualitatively — explanations are described as consistent because the same one or two dominant features appear — without a rank-correlation statistic that would expose disagreement further down the ordering. This study measures what those accounts assume.

Materials and Methods

Dataset

The Cleveland subset of the UCI Heart Disease database was used, obtained from a public mirror of the processed.cleveland file. The raw file contains 303 records; one exact duplicate was removed, leaving 302 records with 13 predictors and a binary target. The cohort is close to balanced: 164 cases with disease present and 138 without, a prevalence of 54.3%. Table 1 reports the descriptive statistics computed on the de-duplicated cohort.

Table 1. Descriptive statistics of the analysis cohort ($n = 302$)

Predictor	Variable	Mean	SD	Range
Age (years)	age	54.42	9.05	29–77
Sex (1 = male)	sex	0.68	0.47	0–1
Chest pain type	cp	0.96	1.03	0–3
Resting blood pressure (mm Hg)	trestbps	131.60	17.56	94–200
Serum cholesterol (mg/dL)	chol	246.50	51.75	126–564
Fasting glucose > 120 mg/dL	fbs	0.15	0.36	0–1
Resting ECG result	restecg	0.53	0.53	0–2
Maximum heart rate achieved	thalach	149.57	22.90	71–202
Exercise-induced angina	exang	0.33	0.47	0–1
ST depression (exercise vs rest)	oldpeak	1.04	1.16	0–6.2
Slope of peak exercise ST segment	slope	1.40	0.62	0–2
Major vessels coloured by fluoroscopy	ca	0.72	1.01	0–4
Thalassemia category	thal	2.31	0.61	0–3

Note. Categorical predictors are reported as numeric codes as supplied in the source file; means for binary variables represent prevalence.

Models

Four classifiers spanning the interpretability spectrum were evaluated. An L2-regularised logistic regression on standardised inputs serves as the inherently interpretable reference. A random forest (500 trees, minimum leaf size 2) and a gradient-boosted tree ensemble (400 rounds, depth 3, learning rate 0.05, subsample and column-sample 0.9, L2 penalty 1.0) represent the black-box models routinely reported in this literature [15-19].

The fourth model was implemented specifically so that learned attention could be evaluated on tabular clinical data. It applies a single additive attention layer over the standardised feature vector followed by a linear read-out:

$$\begin{aligned} a &= \text{softmax}(A x + c) && \text{attention weights over the 13 predictors} \\ z &= w \cdot (a \odot x) + b && \text{gated linear read-out} \\ p &= \sigma(z) && \text{predicted probability of disease} \end{aligned}$$

Parameters were fitted by L-BFGS on the penalised log-loss (L2 coefficient 0.01, up to 600 iterations). The attention vector a is a per-patient probability distribution over predictors and is read directly as that patient's explanation, requiring no post-hoc computation.

Evaluation protocol

Predictive performance was estimated by stratified ten-fold cross-validation at a fixed seed, reporting AUC, accuracy, F1, sensitivity, specificity and the Brier score as mean $\pm$ standard deviation across folds. Receiver-operating-characteristic and calibration curves were constructed from out-of-fold predictions so that every point is an out-of-sample estimate. For the explanation study the data was split once into a 70% training partition ($n = 211$) and a 30% held-out partition ($n = 91$); all attributions were computed on the held-out partition only. Figure 1 summarises the pipeline.

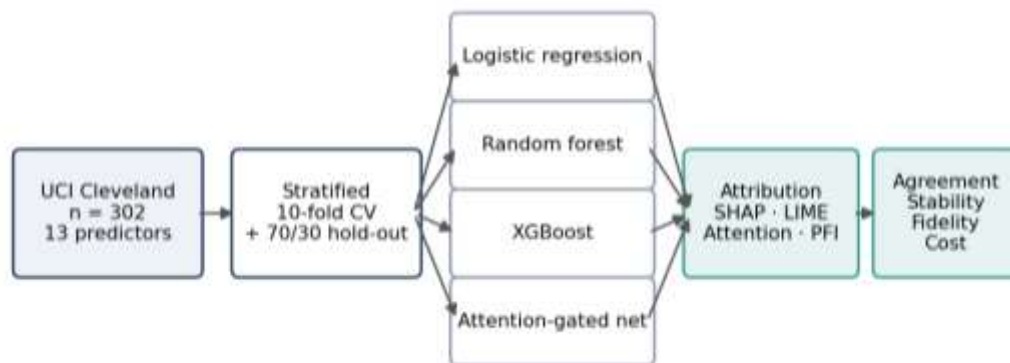


Figure 1. Experimental pipeline. Four classifiers are trained under identical partitioning; four attribution methods are then evaluated on the same held-out patients along four axes.

Attribution methods and comparison metrics

SHAP values were computed with TreeSHAP on the gradient-boosted model. LIME explanations used the tabular explainer with 1,000 perturbation samples and discretised continuous features. Attention weights were read from the attention-gated network. Permutation importance used 30

repeats scored on AUC. Global importance for each method is the mean absolute attribution across held-out patients, normalised to sum to one for comparability.

Four comparison metrics were computed:

- Agreement. Spearman ρ and Kendall τ between the global importance vectors of each method pair, plus the overlap of the top-five feature sets.
- Stability. For six held-out patients, each explanation was regenerated 25 times with the model, data and instance held constant. Instability is summarised as the mean coefficient of variation of the attributions and as the mean pairwise Jaccard similarity of the top-five feature sets across runs.
- Fidelity. A per-instance deletion curve: features are replaced by the training median in descending order of that patient's own attribution magnitude, and the model's confidence in its originally predicted class is tracked. A faithful attribution removes decision-relevant evidence first, so a lower area under the deletion curve (AUDC) indicates higher fidelity. A random-order baseline averaged over ten permutations provides the reference.
- Cost. Mean wall-clock time to produce one patient's explanation, measured on identical hardware within a single process.

The complete pipeline is deterministic at seed 42 with the sole exception of the stability experiment, where LIME's random state is deliberately left unfixed since that stochasticity is the object of measurement [20-27].

Results

Predictive performance

Discrimination was similar across all four model families (Table 2, Figures 2 and 3). The random forest achieved the highest cross-validated AUC at 0.909 ± 0.059 , followed by logistic regression at 0.899 ± 0.060 , gradient boosting at 0.885 ± 0.054 and the attention-gated network at 0.876 ± 0.059 . The 0.010 AUC separating the best ensemble from the logistic regression is an order of magnitude smaller than the between-fold standard deviation of either, so on this cohort the added complexity purchases no measurable discrimination.

The models differ more in operating characteristics than in ranking ability. Logistic regression was the most sensitive (0.897) and the attention network the most specific (0.805), a trade-off that matters more for a screening application than the AUC difference does. Calibration (Figure 4) separates the models more clearly than discrimination: logistic regression and random forest achieved Brier scores of 0.125, gradient boosting 0.153 and the attention network 0.179, the last being visibly over-confident at the extremes.

Table 2. Cross-validated performance (stratified 10-fold, mean $\pm$ SD)

Model	AUC	Accuracy	F1	Sensitivity	Specificity	Brier
Logistic regression	0.899 ± 0.060	0.831 ± 0.064	0.852 ± 0.056	0.897 ± 0.065	0.754 ± 0.090	0.125 ± 0.041
Random forest	0.909 ± 0.059	0.838 ± 0.056	0.855 ± 0.052	0.885 ± 0.084	0.784 ± 0.094	0.125 ± 0.031

Model	AUC	Accuracy	F1	Sensitivity	Specificity	Brier
XGBoost	0.885 ± 0.054	0.794 ± 0.068	0.813 ± 0.066	0.831 ± 0.111	0.754 ± 0.107	0.153 ± 0.051
Attention-gated network	0.876 ± 0.059	0.801 ± 0.067	0.813 ± 0.064	0.799 ± 0.077	0.805 ± 0.111	0.179 ± 0.061

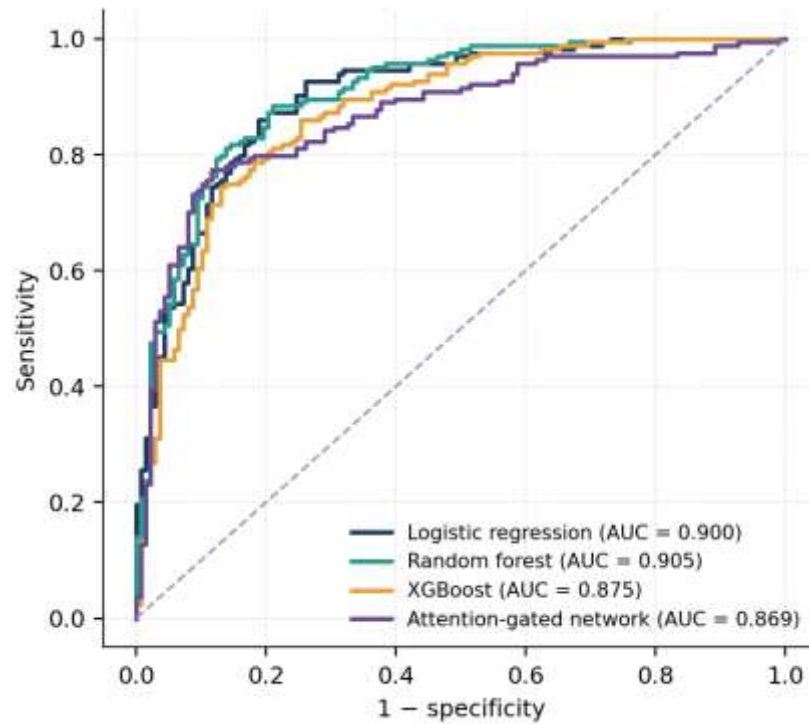


Figure 2. Receiver-operating-characteristic curves constructed from out-of-fold predictions across all ten folds, so every point is an out-of-sample estimate.

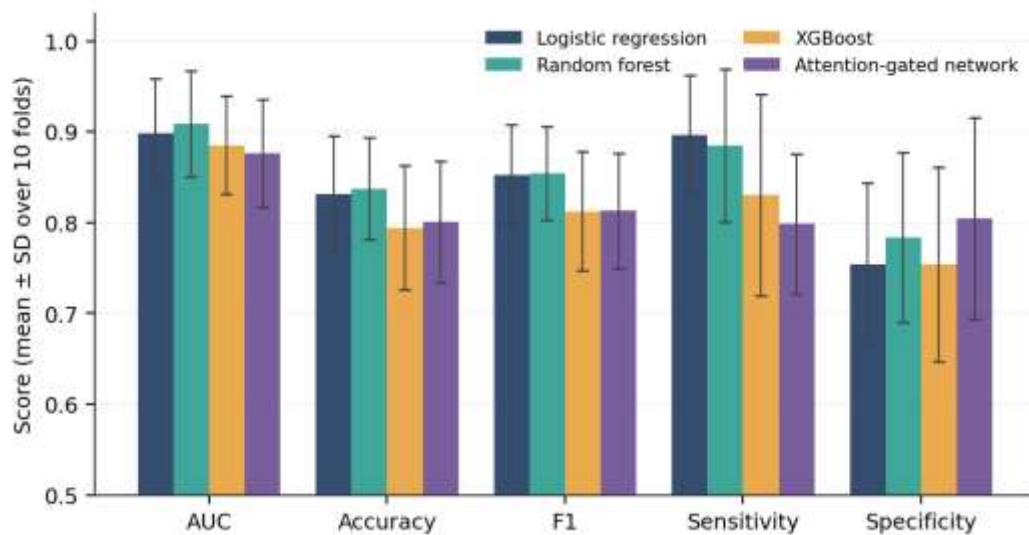


Figure 3. Cross-validated performance across five metrics. Error bars are one standard deviation across the ten folds; the overlap between models on every metric is substantial.

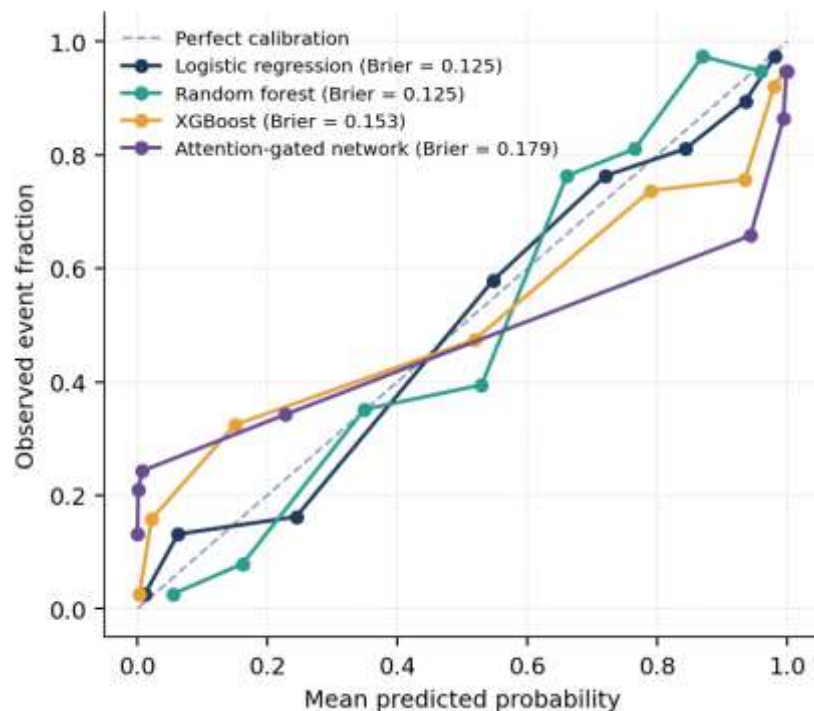


Figure 4. Calibration curves with eight quantile bins. The attention-gated network deviates most from the diagonal, consistent with its higher Brier score.

Global feature importance and method disagreement

The four attribution methods produce visibly different importance profiles from the same model (Figures 5 and 6). All four place chest pain type, thalassemia category and the number of major vessels coloured by fluoroscopy in their top five, which is reassuring and clinically coherent. Below that the orderings separate sharply.

The clearest disagreement concerns serum cholesterol. SHAP ranks it third of thirteen predictors; LIME ranks it fifth; the attention network ranks it tenth; permutation importance ranks it last. This is not a marginal reordering — it is the difference between a clinician being told that cholesterol drove the prediction and being told it was irrelevant. The explanation lies in what each method measures. SHAP quantifies how much the model's output moves with the feature, and the gradient-boosted model does use cholesterol. Permutation importance quantifies how much discriminative performance is lost when the feature is destroyed, and here the model recovers almost entirely through correlated predictors. Both statements are true; they answer different questions, and only one of them answers the clinician's [28-34].

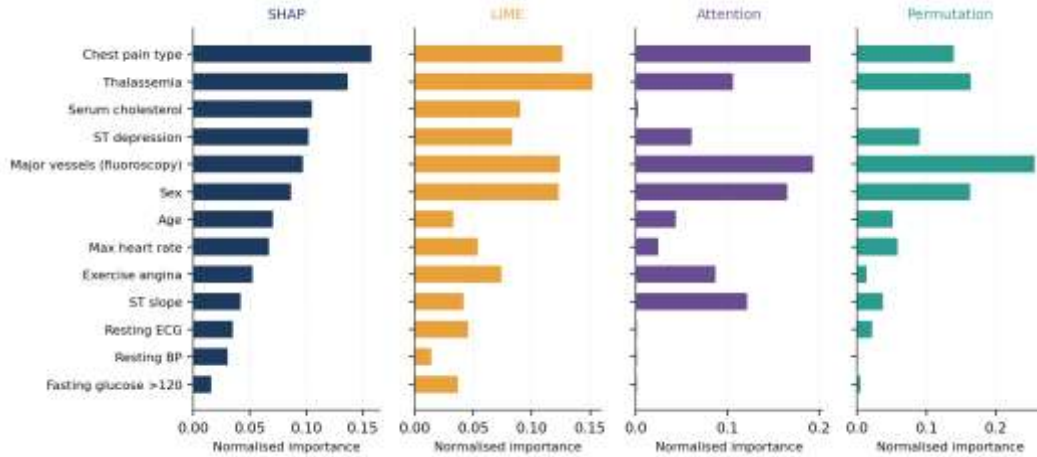


Figure 5. Normalised global feature importance by method, with features ordered by SHAP importance to make divergence in the other panels directly visible.

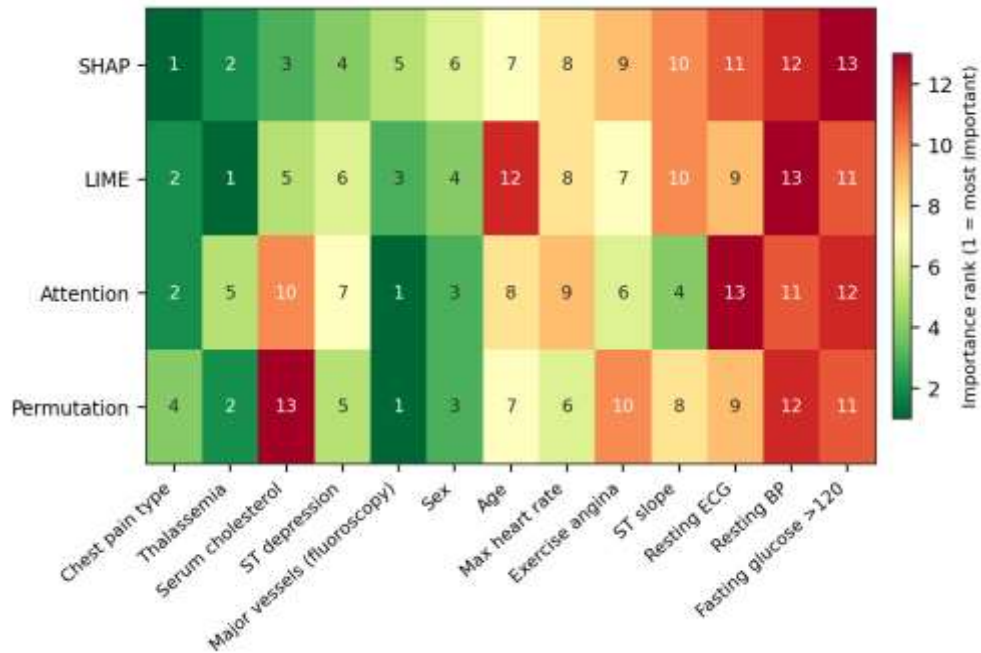


Figure 6. Importance rank by method (1 = most important). Serum cholesterol moves from rank 3 under SHAP to rank 13 under permutation importance.

Table 3. Top five predictors identified by each attribution method

Rank	SHAP	LIME	Attention	Permutation
1	Chest pain type	Thalassemia	Major vessels	Major vessels
2	Thalassemia	Chest pain type	Chest pain type	Thalassemia
3	Serum cholesterol	Major vessels	Sex	Sex
4	ST depression	Sex	ST slope	Chest pain type
5	Major vessels	Serum cholesterol	Thalassemia	ST depression

Note. Three predictors — chest pain type, thalassemia and major vessels — appear in all four top-five lists; the union of the four lists spans seven distinct predictors.

Quantifying the divergence (Figure 7, Table 4), SHAP and LIME agree most closely ($\rho = 0.846$, $\tau = 0.667$), which is expected given that both are local perturbation-based methods applied to the same model. SHAP and permutation importance agree least ($\rho = 0.582$, $\tau = 0.487$). Attention correlates more strongly with permutation importance ($\rho = 0.764$) than with either post-hoc method, notwithstanding that attention comes from a different model altogether. No pair exceeds $\rho = 0.85$, and the mean off-diagonal correlation is 0.689.

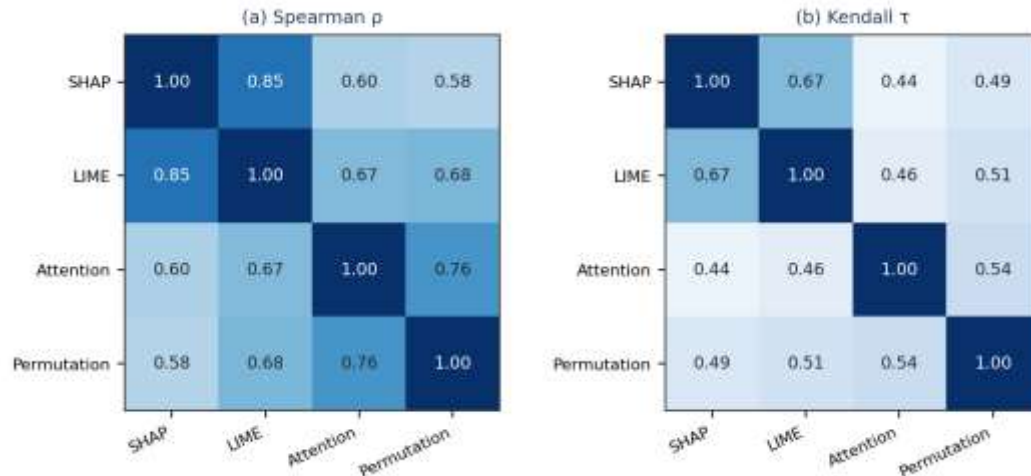


Figure 7. Pairwise rank agreement between attribution methods: (a) Spearman ρ , (b) Kendall τ . No pair achieves the agreement that would justify treating the methods as interchangeable.

Table 4. Pairwise rank agreement (Spearman ρ above, Kendall τ below)

	SHAP	LIME	Attention	Permutation
SHAP	—	0.846	0.599	0.582
LIME	0.667	—	0.665	0.676
Attention	0.436	0.462	—	0.764
Permutation	0.487	0.513	0.538	—

Stability under repeated explanation

Regenerating each explanation 25 times for six unchanged patients with an unchanged model isolates the stochasticity of the explanation algorithm itself. SHAP and attention were exactly reproducible: maximum absolute deviation across runs was 0.0 to machine precision for both, as their computations are deterministic given the model.

LIME was not. Attribution magnitudes varied with a mean coefficient of variation of 0.279, ranging from 0.236 to 0.374 across the six patients — meaning a typical LIME attribution fluctuates by roughly 28% of its own magnitude between identical runs. The top-five feature set was more robust than the magnitudes, with a mean pairwise Jaccard similarity of 0.799, but for one patient it fell to 0.688, implying that the fifth-ranked risk factor shown to a clinician would change depending on when the explanation happened to be generated (Figure 8).

This is a design property of LIME rather than a defect in the implementation: the surrogate is fit to randomly drawn perturbations, so a different draw gives a different surrogate. It can be mitigated by fixing a random seed or by increasing the sample count, but both remedies have costs — a fixed

seed produces reproducibility without stability, and more samples increase an already substantial computational burden.

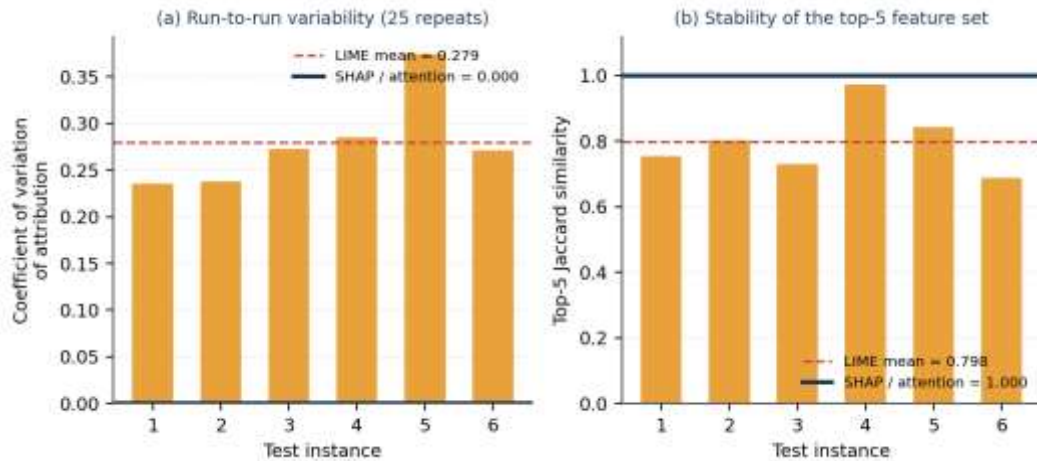


Figure 8. Run-to-run stability over 25 repeated explanations of six unchanged patients. (a) Coefficient of variation of attribution magnitude. (b) Jaccard similarity of the top-five feature set. SHAP and attention are exactly reproducible and appear as reference lines.

Fidelity to model behaviour

All four methods outperformed the random-order baseline on the deletion test, confirming that each carries genuine signal about the model's decision process (Figure 9). SHAP was most faithful (AUDC 0.662), marginally ahead of LIME (0.665); attention (0.691) and permutation importance (0.694) trailed, against a random baseline of 0.739.

The ordering is interpretable. SHAP and LIME produce per-patient attributions, so the deletion order is tailored to each case; permutation importance is global by construction and applies the same order to every patient, which is a handicap on a per-instance test. Attention is nominally local but, as Section 4.6 shows, concentrates its mass so narrowly that after the first two removals its ordering carries little information. The gap between the best method and the random baseline — 0.077 AUDC — is meaningful but not dramatic, a reminder that attribution methods capture the model's reasoning approximately rather than exactly [35-39].

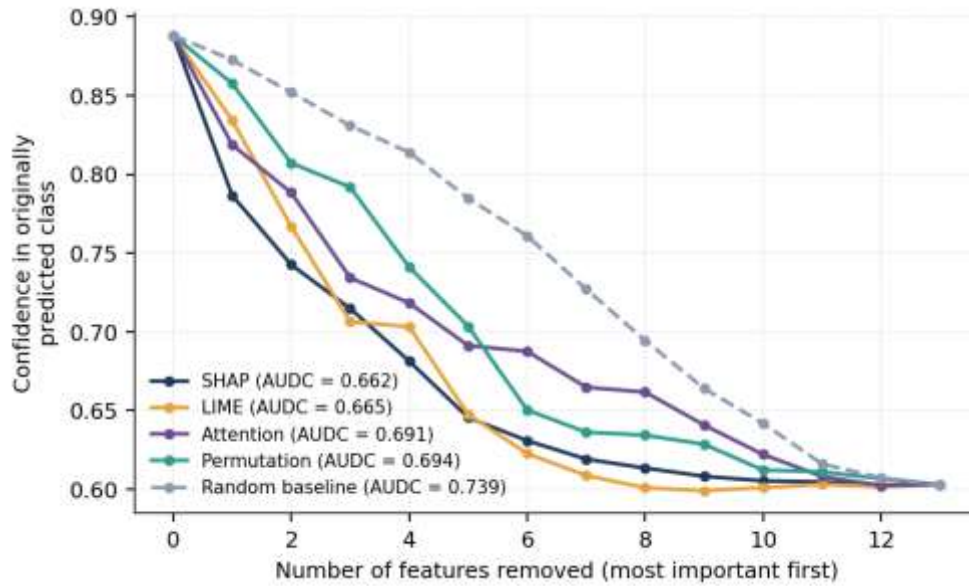


Figure 9. Per-instance deletion curves. Features are removed in descending order of each patient's own attribution magnitude; a steeper fall indicates a more faithful attribution. AUDC values are given in the legend, where lower is better.

Computational cost

Cost per explained patient spans nearly four orders of magnitude (Figure 10). Reading attention weights took 0.002 ms because the explanation is a by-product of the forward pass. TreeSHAP required 0.658 ms. Permutation importance took 8.505 ms per patient at 30 repeats, and LIME 13.48 ms — approximately 20 times slower than SHAP and 6,700 times slower than attention.

In a batch research setting these differences are immaterial. In an interactive decision-support tool they are not: explaining a 500-patient ward round costs about 7 seconds with LIME against 0.3 seconds with SHAP and one millisecond with attention. Combined with the stability results, this makes LIME difficult to recommend for the tabular clinical setting studied here — it is simultaneously the slowest method and the only irreproducible one, while being no more faithful than SHAP.

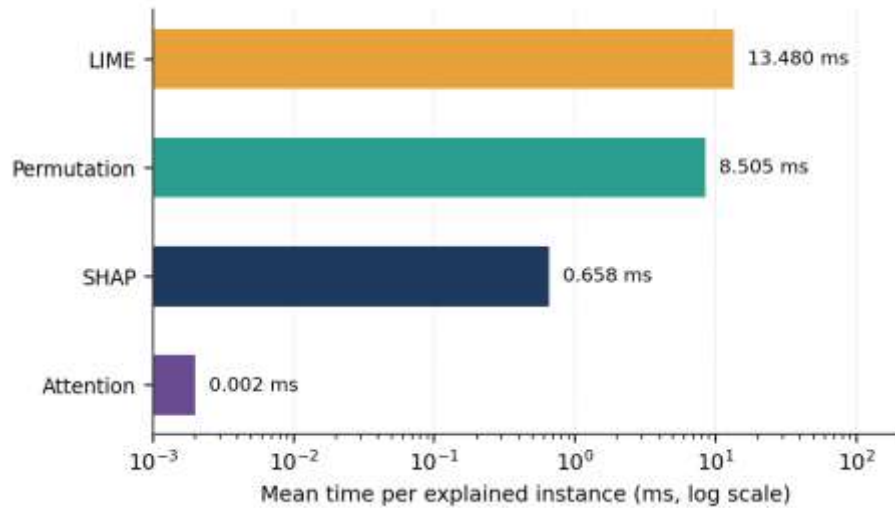


Figure 10. Mean wall-clock time per explained patient, log scale. Attention is essentially free because it is produced during inference.

Table 5. Summary comparison of the four attribution methods

Property	SHAP	LIME	Attention	Permutation
Scope	Local + global	Local	Local	Global only
Model access required	Tree structure	Predictions only	Built into model	Predictions only
Fidelity (AUDC, lower better)	0.662	0.665	0.691	0.694
Reproducible across runs	Yes (dev. 0.0)	No (CV 0.279)	Yes (dev. 0.0)	Approximately
Top-5 stability (Jaccard)	1.000	0.799	1.000	—
Cost per patient (ms)	0.658	13.480	0.002	8.505
Main limitation	Needs a compatible model	Slow and unstable	Collapses onto few features	Blind to correlated features

Local case study and attention collapse

A single held-out patient illustrates how the methods differ at the point of care (Figure 11). The patient is a 37-year-old woman with atypical chest pain, cholesterol 215 mg/dL, maximum heart rate 170, no exercise-induced angina, no ST depression and no vessels coloured by fluoroscopy. The gradient-boosted model predicted disease with probability 0.9995 against a base value of 0.166, and the true label confirms disease present.

SHAP distributes the explanation across several predictors: chest pain type contributes +1.104 in log-odds, serum cholesterol +0.973, thalassemia +0.892, sex +0.880 and maximum heart rate +0.851. LIME agrees on direction but with a different ordering, ranking sex (+0.215) and thalassemia (+0.221) above chest pain type (+0.162) [40-45].

The attention network behaved quite differently. It assigned 96.0% of its total attention mass to sex, 2.6% to chest pain type and 1.2% to ST slope, leaving under 0.2% distributed across the remaining ten predictors. As a clinical explanation this is close to useless — it states that the prediction was about the patient being female and almost nothing else. This attention collapse was consistent across held-out patients and explains the method's weak deletion-curve performance in Section 4.4: an attribution vector concentrated on one feature provides no meaningful ordering for the remaining twelve. It is a concrete instance of the concern that attention weights are not automatically explanations [13].

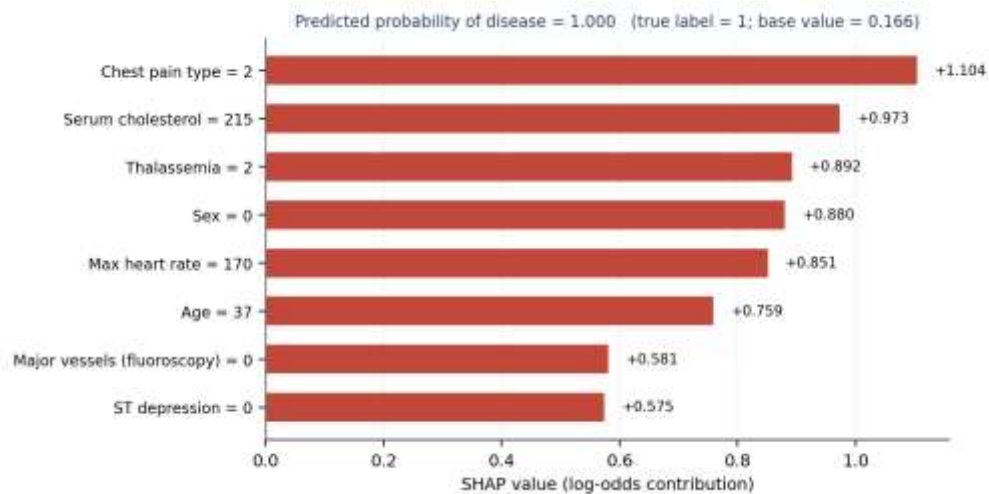


Figure 11. SHAP attributions for a single held-out patient, showing the eight predictors with the largest absolute contribution. Red bars increase and blue bars decrease the predicted log-odds of disease.

Table 6. Attributions for the case-study patient across three methods

Predictor	Value	SHAP (log-odds)	LIME (weight)	Attention (share)
Chest pain type	2	+1.104	+0.162	2.6%
Serum cholesterol	215 mg/dL	+0.973	+0.067	0.0%
Thalassemia	2	+0.892	+0.221	0.0%
Sex (female)	0	+0.880	+0.215	96.0%
Maximum heart rate	170	+0.851	+0.126	0.0%
Age	37	+0.759	+0.043	0.0%
Major vessels	0	+0.581	+0.196	0.1%
ST depression	0.0	+0.575	+0.119	0.0%
ST slope	2	+0.344	+0.063	1.2%
Exercise angina	0	+0.326	+0.108	0.0%

Discussion

Explanation method is a clinical decision, not an implementation detail

The central finding is that four defensible attribution methods, applied to the same model and the same patients, disagree enough to change the clinical narrative. Mean pairwise rank correlation of 0.689 means the methods concur on the strongest signals and diverge elsewhere; the cholesterol

case, moving from rank 3 to rank 13, shows how large that divergence can be. A clinician shown a SHAP explanation and a clinician shown a permutation-importance explanation of the identical prediction would form materially different views of what mattered.

The practical implication is that the choice of explanation method should be documented and justified in the same way as the choice of model, and that a single attribution method should not be presented as the explanation of a prediction. Where two methods that answer different questions disagree, both results should be shown rather than the more convenient one selected. The applied literature does not currently work this way; method selection is often justified on convenience grounds, as in the study that chose permutation importance over SHAP because SHAP's cost was prohibitive for a multi-model design [8] — a defensible engineering decision that nonetheless changes the answer given to the clinician.

Attention is not explanation, at least not here

The attention-gated network produced the most concentrated and least useful explanations in this study. Assigning 96% of attention mass to one predictor is not interpretability; it is a degenerate solution that the training objective had no reason to avoid, since nothing in the loss rewards a well-distributed attention vector. Attention was also the least faithful of the local methods on the deletion test. This is a single architecture on one tabular dataset and does not generalise to attention in transformer models over sequences or images, but it is a direct empirical illustration of Jain and Wallace's argument [13], and it suggests that attention weights should be validated against an independent attribution method before being presented to clinicians as reasoning.

The interpretable baseline was not beaten

A logistic regression achieved 0.899 AUC against the random forest's 0.909 — a difference of 0.010 against between-fold standard deviations near 0.059 — while matching the best Brier score at 0.125. The logistic model requires no attribution method at all: its coefficients are the explanation, they are exactly reproducible, they cost nothing to compute and they cannot disagree with themselves.

This is the empirical situation Rudin describes [3]. Where an interpretable model performs equivalently, layering a black box and a post-hoc explainer on top introduces every problem documented in this study — method disagreement, run-to-run instability, imperfect fidelity, computational cost — in exchange for no measurable benefit. The finding should not be over-generalised: this is a small, clean, low-dimensional tabular cohort, exactly the regime in which linear models are competitive, and it says nothing about imaging or high-dimensional omics. But it does suggest that the interpretable baseline deserves to be reported alongside the black box in clinical work far more often than it is.

Practical guidance

For tabular clinical decision support of the kind studied here, the results support four recommendations:

- Fit an interpretable model first and report it. If it performs comparably, the explainability problem largely dissolves.

- Where a black box is genuinely warranted, prefer SHAP for tree ensembles: it was the most faithful method, exactly reproducible, and 20 times faster than LIME.
- Report at least two attribution methods and disclose their rank correlation. Agreement is evidence; silent selection of a single method is not.
- Treat attention weights as a hypothesis about model reasoning that requires external validation, not as an explanation in themselves.

Limitations

This study is deliberately narrow, and four constraints bound its conclusions. The cohort is small ($n = 302$), single-centre, decades old and close to class-balanced; contemporary clinical data is larger, messier and far more imbalanced, and both model rankings and explanation behaviour may differ there. No external validation cohort was available, so the generalisability of the specific feature rankings is untested.

Second, the attention architecture is a single additive-attention design chosen for transparency rather than performance. A different attention formulation, or an entropy penalty encouraging distributed weights, might avoid the collapse observed here, and the negative finding should be read as being about this architecture rather than about attention in general.

Third, deletion-curve fidelity is one of several competing definitions of faithfulness, and replacing features with the training median moves instances off the data manifold, which can produce predictions the model was never trained to make. Insertion curves, conditional expectation imputation and axiomatic checks would give a fuller picture.

Fourth, and most importantly, no clinician evaluated any of these explanations. Fidelity, stability and agreement are properties of the algorithm; usefulness is a property of the interaction between an explanation and a human decision-maker. A stable and faithful explanation that clinicians cannot act on has not solved the problem this literature exists to solve.

Future Work

Four extensions follow directly. External validation on a larger, more recent and imbalanced cardiovascular cohort would establish whether the disagreement patterns reported here persist when class prevalence and dimensionality change. A clinician-in-the-loop study measuring decision quality — not satisfaction — under explanations from different methods would test whether the measured differences translate into different clinical actions. Attention architectures trained with an explicit entropy or sparsity-balance penalty would establish whether collapse is avoidable by design rather than accidental. Finally, an ensemble attribution that aggregates several methods and reports their disagreement as an explicit uncertainty band, rather than presenting one method's output as definitive, appears a more honest interface than any single method's plot [46-47].

applied to identical models and patients agreed only moderately (mean pairwise Spearman $\rho = 0.689$), with one predictor moving from third to last in importance depending on the method chosen. LIME was the slowest method at 13.48 ms per patient and the only irreproducible one, with attributions varying by a mean coefficient of variation of 0.279 across identical repeated runs. SHAP was the most faithful on a per-instance deletion test (AUDC 0.662) and exactly reproducible.

Learned attention was essentially free to compute but collapsed 96% of its mass onto a single predictor and was the least faithful local method.

Underlying all of this, a regularised logistic regression matched every black-box model on discrimination and equalled the best on calibration, requiring no attribution method whatsoever. Explanation methods are approximations that disagree with one another, and treating any single one as the reason for a clinical prediction overstates what it can deliver. Where an interpretable model suffices, it remains the strongest available guarantee — and where it does not, the explanation method should be selected, justified and reported with the same rigour as the model itself.

References

- [1] Roth, G. A., Mensah, G. A., Johnson, C. O., Addolorato, G., Ammirati, E., Baddour, L. M., ... Fuster, V. (2020). Global burden of cardiovascular diseases and risk factors, 1990–2019. *Journal of the American College of Cardiology*, 76(25), 2982–3021.
- [2] Virani, S. S., Alonso, A., Aparicio, H. J., Benjamin, E. J., Bittencourt, M. S., Callaway, C. W., ... Tsao, C. W. (2021). Heart disease and stroke statistics — 2021 update. *Circulation*, 143(8), e254–e743.
- [3] Ahmed, K. R., Sufian, M. A., Faruq, O., Kowser, S., Shammah, R. S., & Ahmmed, M. R. (2025, July). Impact of IT product innovation on competitive advantage in the digital economy. In *2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)* (pp. 1-6). IEEE.
- [4] 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.
- [5] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4768–4777.
- [6] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.
- [7] Siddiqui, M. I. H., Rahman, M. S., Kabir, A. A., Mahmud, F. U., Rashid, S. U., & Shammah, R. S. (2023). Comparative analysis of explainable machine learning models for cancer classification using cytological features. *Journal of Medical and Health Studies*, 4(5), 110–150.
- [8] Enhanced cardiovascular disease risk prediction using explainable machine learning and data balancing. (2026). *Discover Computing*.
- [9] Shahiduzzaman, M., Sourav, M. S. A., Kabtia, M., Shammah, R. S., Islam, M. A., Chakraborty, P., ... & Adnan, M. (2026). Attention-guided quantum-inspired extreme learning machine with grey wolf optimization for rare-attack detection in IIoT networks. *Discover Artificial Intelligence*.
- [10] Ensemble learning approach with explainable AI for improved heart disease prediction. (2026). *Frontiers in Pharmacology*.

- [11] Dual explainability framework for heart disease prediction using LIME and permutation feature importance. (2025). *Discover Applied Sciences*.
- [12] Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., & Lakkaraju, H. (2022). *The disagreement problem in explainable machine learning: A practitioner's perspective*. arXiv. <https://arxiv.org/abs/2202.01602>
- [13] Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 180–186.
- [14] An explainable machine learning framework for cardiovascular risk prediction using structured health data. (2026). *Frontiers in Artificial Intelligence*.
- [15] Explainable soft-voting classifier for heart disease prediction using SHAP and LIME. (2026). *Discover Computing*.
- [16] Jain, S., & Wallace, B. C. (2019). Attention is not explanation. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 3543–3556.
- [17] Wiegrefe, S., & Pinter, Y. (2019). Attention is not not explanation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 11–20.
- [18] Detrano, R., János, A., Steinbrunn, W., Pfisterer, M., Schmid, J.-J., Sandhu, S., Guppy, K. H., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *American Journal of Cardiology*, 64(5), 304–310.
- [19] Dua, D., & Graff, C. (2019). UCI Machine Learning Repository: Heart Disease Data Set. University of California, Irvine, School of Information and Computer Sciences.
- [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.
- [21] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- [22] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
- [23] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- [24] Karim, N. B., Tolosana, D., & Simko, R. (2025b). Responsible and Legally Compliant Enterprise Artificial Intelligence in Small-Business Lending: Governance, Explainability, and Fair-Lending Safeguards for Early-Intervention Recommendation Engines. *International Journal of Modern Computing*, 8(1), 55-66.

- [25] Rahman, M. S., Islam, C. S., & Mudryi, M. (2025). Interoperability Standards for Real-Time Fraud-Signal Exchange Across US Financial Institutions: Design Principles and a Reference Implementation. *The Computertech*, 47-60.
- [26] Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.).
- [27] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3.
- [28] Rashid, S. U., & Sattiraju, S. R. (2025b). Multi-Institution Validation of Drift-Resilient Early-Warning Models for SME Loan Distress and Transaction Fraud: Evidence from US Community Financial Institution Pilots. *International Journal of Modern Computing*, 8(1), 40-54.
- [29] Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games II*, 307–317.
- [30] Samek, W., Binder, A., Montavon, G., Lapuschkin, S., & Müller, K.-R. (2017). Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11), 2660–2673.
- [31] Alvarez-Melis, D., & Jaakkola, T. S. (2018). *On the robustness of interpretability methods*. arXiv. <https://arxiv.org/abs/1806.08049>
- [32] Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11), e745–e750.
- [33] Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20, 310.
- [34] 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.
- [35] Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31, 841–887.
- [36] Saeed, W., & Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263, 110273.
- [37] Explainable hybrid deep learning framework for cardiovascular disease prediction and clinical decision support. (2026). PubMed 42411231.
- [38] Rahman, M. S., Islam, C. S., Mudryi, M., & Cakmak, S. (2025). Concept-Drift-Resilient Ensemble Learning for Financial Transaction Fraud Detection Under Evolving Criminal Typologies. *International Journal of Acta Informatica*, 4(1), 52-67.

- [39] A multi-tier explainable AI architecture for the prediction of cardiovascular diseases. (2026). *Proceedings of the 9th International Conference on Information and Computer Technologies*.
- [40] An explainable artificial intelligence framework for ischemic heart disease prediction using enhanced squirrel search feature selection. (2026). *Scientific Reports*.
- [41] Potharaju, B., & Ravipudi, J. (2025). Automated Data Validation and Resilient Infrastructure Design for Enterprise Regulatory Reporting Platforms. *International Journal of Computer Science and Engineering Innovations*, 1(1), 46-64.
- [42] Predicting cardiovascular risk with hybrid ensemble learning and explainable AI. (2025). *Scientific Reports*.
- [43] Rashid, S. U., & Sattiraju, S. R. (2025a). Data-Driven Early-Warning Indicators of Small-Business Financial Distress: Integrating Operational, Transactional, and Credit Data for Decision Support in SMEs. *The Metascience*, 3(4), 53-65.
- [44] Karim, N. B., & Kovac, T. (2025). Cloud-Based Big-Data Architectures for AI-Driven Financial Planning and Analysis and Early Risk Signaling in Resource-Constrained Community Financial Institutions. *The Metascience*, 3(3), 29-40.
- [45] Sokol, K., & Flach, P. (2020). Explainability fact sheets: A framework for systematic assessment of explainable approaches. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 56–67.
- [46] Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16, 1–85.
- [47] Potharaju, B., & Ravipudi, J. (2023). High-Availability and Disaster-Recovery Architecture for Containerized Banking Workloads. *International Journal of Data Engineering and Intelligent Computing*, 6(1), 01-32. Potharaju & Ravipudi, 2023).