International Journal of Scientific Engineering and Research (IJSER)
ISSN (Online): 2347-3878
SJIF (2025): 8.036

An Interpretable Machine Learning Framework with Cross-Model Feature Stability Analysis for Heart Disease Prediction

Rekha P[1], Dr. Malliga P[2]

[1]Post-Graduate Student, Department of Computer Science and Engineering, National Institute of Technical Teachers Training and Research (NITTTR), Chennai, India

[2]Professor, Department of Computer Science and Engineering, National Institute of Technical Teachers Training and Research (NITTTR), Chennai, India

Abstract: ***Machine learning models for heart disease prediction are typically compared on discrimination alone, and the feature importances of a single selected model are then reported as though they described the data rather than the model. This paper develops an interpretable framework in which cross-model agreement is treated as a measurable quantity rather than an assumption. Logistic Regression, Random Forest, and Extreme Gradient Boosting (XGBoost) are trained on the Cleveland heart disease dataset, and SHapley Additive exPlanations (SHAP) are computed for every model at both the cohort and the individual-patient level. A Feature Stability Score (FSS) is used to combine the three attribution profiles into a single ranking. We show analytically and empirically that the variance-based form of this score, FSS = 1 − σ², is not scale-invariant and is therefore confounded with importance itself: across the thirteen predictors its rank correlation with mean importance is ρ = −0.973, so the least important feature is scored as the most stable, and the resulting composite ranking correlates with raw importance at ρ = +0.995 - the stability term changes essentially nothing. We propose a scale-invariant reformulation based on the coefficient of variation of per-model normalised attributions, under which stability becomes an independent signal (ρ = +0.703) and the ranking changes materially: oldpeak rises from seventh to fourth while sex, chol, and age fall, the latter because Logistic Regression and XGBoost disagree about it by a factor of thirty. Random Forest attains the best discrimination (accuracy 0.8361, ROC-AUC 0.9188). At the individual level the three models disagree sharply on a representative patient, returning probabilities of 0.4987, 0.6224, and 0.1995 and splitting on the resulting decision, while the patient-level feature ranking diverges from the cohort ranking (ρ = 0.698, with eleven of thirteen positions changed). These results indicate that a single-model explanation of a heart disease classifier is not sufficient evidence about the underlying predictors.***

Keywords: Heart disease prediction, explainable artificial intelligence, SHAP, feature stability, cross-model agreement, model disagreement, clinical decision support.

1. Introduction

Cardiovascular disease remains the leading cause of death worldwide, accounting for roughly one third of all global deaths [1]. A substantial share of that burden is attributable to coronary artery disease, whose early stages are frequently asymptomatic. Because the clinical measurements relevant to risk- resting blood pressure, cholesterol, exercise-induced angina, ST depression, and the results of a stress test- are routinely collected and inexpensive, automated risk stratification from structured clinical records has been an active area of applied machine learning for three decades [2], [3].

The dominant methodological template in this literature is well established: train several classifiers on a tabular dataset, compare them on accuracy and the area under the receiver operating characteristic curve (ROC-AUC), select the best, and report its feature importances as an account of which clinical variables drive risk. Recent work has augmented the final step with post-hoc attribution methods, most commonly SHapley Additive exPlanations (SHAP) [4], [5], which decompose an individual prediction into additive per-feature contributions grounded in cooperative game theory [6].

This template contains an unexamined inferential step. The feature importances reported at the end belong to *one model*. They are presented, implicitly or explicitly, as a statement about the data- as though the analysis had discovered which clinical variables matter- when they are in fact a joint property of the data and of a modelling choice that was made on the basis of a fractional difference in accuracy. If a second model of comparable accuracy would have produced a different ranking, then the reported explanation is not a finding but an artefact of model selection.

Whether this matters is an empirical question, and it is answerable. If several models are trained and explained, the degree to which they agree can be measured, and features can be separated into those whose importance is robust to the modelling choice and those whose importance is not. A stability measure of this kind is the natural bridge between multi-model comparison and a single reportable conclusion.

The present work implements such a framework for heart disease prediction and, in the course of evaluating it, identifies a defect in the natural first formulation of the stability measure. Defining stability as one minus the cross-model variance of the attributions is intuitive, but variance is not scale-invariant:

multiplying all attributions by a constant multiplies the variance by its square. Because important features necessarily have larger attributions than unimportant ones, their cross-model variance is larger for reasons that have nothing to do with disagreement. The measure consequently reports that the weakest predictor in the dataset is the most stable, and the composite ranking it feeds is almost perfectly determined by importance alone. We quantify this effect, show that it is a scale artefact rather than a property of the data, and give a corrected scale-invariant formulation under which stability carries genuine independent information.

Contributions

1) We build a reproducible heart disease prediction pipeline on the Cleveland dataset with three models spanning the linear, bagged, and boosted families, and compute SHAP attributions for all three at both the cohort and the individual-patient level.
2) We show that the variance-based Feature Stability Score is confounded with importance, quantifying the effect at $\rho = -0.973$ between stability and importance and $\rho = +0.995$ between the composite score and raw importance, and we explain the mechanism analytically.
3) We propose a scale-invariant reformulation based on the coefficient of variation of per-model normalised attributions, demonstrate that it decouples stability from importance ($\rho = +0.703$), and report the corrected feature ranking.
4) We show that the cohort-level ranking does not transfer to an individual patient ($\rho = 0.698$; eleven of thirteen positions change), and that the three models disagree sharply enough on that patient to split on the clinical decision.

Section II reviews related work; Section III describes the data, models, and both formulations of the stability measure; Section IV presents results; Sections V and VI discuss and bound them; Section VII concludes.

2. Related Work

1) *Machine Learning for Heart Disease Prediction*

The Cleveland subset of the UCI Heart Disease collection, assembled by Detrano *et al.* [2], [3], is the standard benchmark for this task. Its 303 records and thirteen predictors have been used to evaluate essentially every standard classifier family, and reported accuracies commonly fall between 0.80 and 0.88.

Comparisons across this literature are complicated by inconsistent protocols. Studies differ in whether the outcome is binarised or kept as the original five-level severity scale, whether the split is stratified, whether categorical variables such as chest-pain type and thalassaemia are one-hot encoded or treated as ordinal integers, and whether hyper-parameters are tuned against the test partition. Because the test partition contains only about sixty records under a standard 80/20 split, differences of one or two percentage points in accuracy correspond to a single reclassified patient and are not meaningful. This observation motivates the position taken here: with a dataset of this size, selecting a model on the basis of small accuracy differences and then reporting its explanation as definitive is not well founded.

2) *Attribution Methods and Their Agreement*

SHAP [4] assigns each feature the Shapley value [6] of a cooperative game whose payoff is the model output, and is the unique attribution satisfying local accuracy, missingness, and consistency; TreeSHAP [5] makes the computation tractable for tree ensembles. LIME [7] provides an alternative through local surrogate models, and Štrumbelj and Kononenko [8] developed closely related contribution-based explanations.

The reliability of such explanations is contested. SHAP's standard estimators assume feature independence and can misattribute credit among correlated predictors [9]; several non-equivalent Shapley formulations exist for model explanation [10]; explanations can be adversarially manipulated [11]; and Rudin [12] argues that intrinsically interpretable models should be preferred outright for high-stakes decisions. Ghorbani *et al.* [13] showed that interpretations can be substantially altered by input perturbations that leave the prediction unchanged, and Alvarez-Melis and Jaakkola [14] formalised robustness criteria for interpretability methods.

Most directly relevant is the disagreement problem articulated by Krishna *et al.* [15]: different explanation methods applied to the same model routinely produce different feature rankings, and practitioners resolve the conflict by ad hoc means. Saarela and Jauhiainen [16] report the analogous phenomenon across importance measures. The variant studied here is complementary and, for applied clinical work, more consequential: not different explanation methods on one model, but one explanation method across different models, where the model is precisely what the analyst has chosen arbitrarily.

3) *Stability Measures*

Formal stability measures originate in the feature-selection literature. Kalousis *et al.* [17] introduced similarity-based indices for comparing selected subsets, and Nogueira *et al.* [18] derived a measure with well-defined statistical properties including a known distribution under the null. These works ask how a single algorithm's output varies under resampling of the data.

The question posed here is the transpose: how much does attribution for a fixed dataset vary across model families? Molnar [19] discusses this qualitatively. The contribution of the present paper is not merely to make it quantitative, but to show that the obvious quantification—cross-model variance—fails for a specific and correctable reason, and to supply the correction.

3. Materials and Methods

1) Dataset

Table I: Predictors in the Cleveland Heart Disease Dataset

Name	Description	Type
age	Age in years	Continuous
sex	Sex (1 = male, 0 = female)	Binary
cp	Chest-pain type (0–3)	Categorical
trestbps	Resting blood pressure (mm Hg)	Continuous
chol	Serum cholesterol (mg/dL)	Continuous
fbs	Fasting blood sugar > 120 mg/dL	Binary
restecg	Resting electrocardiographic result	Categorical
thalach	Maximum heart rate achieved	Continuous
exang	Exercise-induced angina	Binary
oldpeak	ST depression induced by exercise	Continuous
slope	Slope of the peak exercise ST segment	Categorical
ca	Number of major vessels (0–4)	Categorical
thal	Thalassaemia result	Categorical
target	Heart disease present (1) or absent (0)	Binary

Experiments use the Cleveland heart disease dataset [2], [3] in its widely distributed binarised form: 303 patient records, thirteen predictors, and a binary target indicating the presence of heart disease. Table I describes the predictors. The cohort contains 138 negative (45.5%) and 165 positive (54.5%) cases, as shown in Fig. 1(a). This near balance is a useful property: unlike many clinical datasets, accuracy is not dominated by a majority class, and the no-skill floor is 54.5%.

The thirteen predictors are heterogeneous in type. Five are continuous or count-valued (age, trestbps, chol, thalach, oldpeak); three are binary (sex, fbs, exang); and five are categorical but encoded as integers (cp, restecg, slope, ca, thal). The integer encoding of the unordered categorical variables is a modelling decision inherited from the standard distribution of the dataset and is retained here for comparability; its consequences are discussed in Section VI.

2) Preprocessing

Clinical tabular datasets frequently encode missing measurements as zero, and a zero value for resting blood pressure, serum cholesterol, or maximum heart rate is physiologically impossible. The three affected columns were therefore screened for zero entries, which would have been reinterpreted as missing and imputed with the column median.

The screen returned no zero values in any of the three columns, so the imputation step was a no-op and the modelled data are identical to the raw data. We report this explicitly rather than omitting it, because the check is a necessary precondition for trusting the descriptive statistics and the attribution analysis that follow, and because its outcome is a property of this particular distribution of the dataset that a reader reproducing the work should be able to confirm. The dataset likewise contains no explicit null values.

3) Experimental Protocol

The cohort was partitioned into 242 training records (80%) and 61 held-out test records (20%) by stratified sampling, preserving the outcome proportions in both partitions; the test partition contains 28 negative and 33 positive cases. A fixed random seed was used throughout so that the partition and all model fits are reproducible.

Features were standardised to zero mean and unit variance for Logistic Regression, with the scaler fitted on the training partition only and applied unchanged to the test partition. The two tree ensembles are invariant to monotone rescaling of individual features and were trained on the raw values. No hyper-parameter search was conducted against the test partition; the values in Section III-D are fixed *a priori* at conventional settings, so the reported metrics are not inflated by selection over the evaluation data.

4) Models

Three model families were selected to span the relevant range of inductive biases.

a) *Logistic Regression:* A linear model in the log-odds of the outcome, fitted by penalised maximum likelihood with the default ℓ_2 penalty and up to 1000 iterations. It is the clinically conventional baseline and, being additive in the log-odds, supplies an intrinsically interpretable reference against which the ensembles' post-hoc explanations can be compared.

b) *Random Forest:* An ensemble of 200 trees [20] of maximum depth 6, each fitted to a bootstrap resample with feature subsampling at every split; the predicted probability is the mean of the per-tree class frequencies. Averaging over decorrelated trees reduces variance and captures interactions unavailable to the linear model.

c) *XGBoost:* A gradient-boosted ensemble [21] of 300 trees with maximum depth 4, learning rate 0.05, and row and column subsampling of 0.8, trained with the log-loss objective. Boosting fits each tree to the residual of the current ensemble, giving it the greatest capacity of the three to fit fine local structure.

Class weighting was not applied, since the outcome is close to balanced.

5) SHAP Attribution

For a model f and instance x, SHAP assigns feature j the value

$$\phi_j(f,\mathbf{x}) = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!\,(|F|-|S|-1)!}{|F|!}\left[f_{S\cup\{j\}}(\mathbf{x}) - f_S(\mathbf{x})\right], \tag{1}$$

where $\mathcal{F}$ is the feature set and f_S the model restricted to subset S. Attributions are additive: $f(x) = \phi_0 + \Sigma_j \phi_j$, with ϕ_0 the expected output over the background distribution. The exact estimator was used for Logistic Regression and TreeSHAP [5] for the ensembles.

Cohort-level importance for model m and feature j is the mean absolute attribution over the N test records,

$$\bar{\phi}_j^{(m)} = \frac{1}{N}\sum_{i=1}^{N}\left|\phi_j\left(f^{(m)}, \mathbf{x}_i\right)\right|, \tag{2}$$

and individual-level importance for a patient x is simply $|\phi_j(f^{(m)}, x)|$.

6) Feature Stability Score: Variance Formulation

Equation (2) produces one importance profile per model. Let $\mathcal{M}$ = {LR, RF, XGB} and write μ_j and σ^2_j for the mean and

variance of $\bar{\phi}^{(m)}_j$ across $m \in \mathcal{M}$. The Feature Stability Score as originally implemented is

$$\mathrm{FSS}_j = 1 - \sigma_j^2, \tag{3}$$

so that perfect cross-model agreement gives $\mathrm{FSS}_j = 1$ and disagreement reduces the score. Importance is min–max normalised across the d features,

$$\tilde{I}_j = \frac{\mu_j - \min_k \mu_k}{\max_k \mu_k - \min_k \mu_k}, \tag{4}$$

and combined with stability into the composite

$$S_j = \alpha \tilde{I}_j + (1-\alpha)\,\mathrm{FSS}_j, \qquad \alpha = 0.6. \tag{5}$$

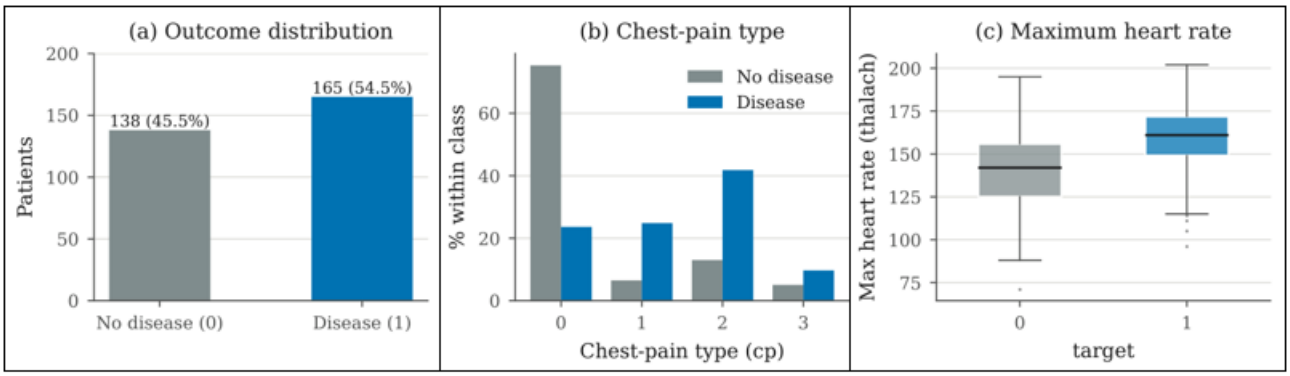

Figure 1: Cohort characteristics. (a) The outcome is close to balanced, so the no-skill accuracy floor is 54.5%. (b) Chest-pain type separates the classes strongly: type 0 (asymptomatic) dominates the negative class while types 1–2 are concentrated in the positive class. (c) Patients with disease attain a higher maximum heart rate under stress in this cohort.

7) Why the Variance Formulation Fails

Equation (3) has two defects, one cosmetic and one substantive.

The cosmetic defect is that it is unbounded below. Variance is not confined to [0, 1], so FSS_j becomes negative whenever $\sigma^2_j > 1$, and the quantity cannot be interpreted as a normalised agreement score. In the present data the attributions are small enough that this does not occur, but the formulation offers no guarantee.

The substantive defect is that variance is not scale-invariant. For any constant c,

$$\mathrm{Var}\left(c\,\bar{\phi}_j^{(\cdot)}\right) = c^2\,\mathrm{Var}\left(\bar{\phi}_j^{(\cdot)}\right), \tag{6}$$

so cross-model variance grows with the squared magnitude of the attributions. Since an important feature has larger attributions in every model than an unimportant one, it incurs a larger variance for reasons unrelated to whether the models agree about it. Writing the attributions of feature j as $\bar{\phi}^{(m)}_j = \mu_j\, r^{(m)}_j$, where $r^{(m)}_j$ is the model's relative weighting, gives $\sigma^2_j = \mu^2_j\,\mathrm{Var}(r^{(\cdot)}_j)$ and hence

$$\mathrm{FSS}_j = 1 - \mu_j^2\,\mathrm{Var}\left(r_j^{(\cdot)}\right). \tag{7}$$

Equation (7) makes the mechanism explicit: unless the relative disagreement $\mathrm{Var}(r^{(\cdot)}_j)$ varies across features by more than the square of their importance, FSS_j is a decreasing function of μ_j and the score measures importance rather than stability. Substituting into (5), the composite becomes $S_j = \alpha\tilde{I}_j + (1-\alpha)(1-\mu^2_j\mathrm{Var}(r^{(\cdot)}_j))$, a sum of one increasing and one decreasing function of importance, whose ordering is determined by whichever term dominates. Section IV-D shows that on these data the first term dominates completely.

This confound is aggravated by a second scale mismatch. Random Forest SHAP values are expressed in probability units, whereas Logistic Regression and XGBoost attributions are in log-odds units and are roughly an order of magnitude larger (Fig. 3). The cross-model variance is therefore driven almost entirely by the two log-odds models, and the Random Forest profile contributes little regardless of whether it agrees.

8) Scale-Invariant Reformulation

Both defects are removed by measuring dispersion relative to magnitude, and by placing the models on a common scale before comparing them.

First, each model's importance profile is normalised to unit sum, which eliminates the unit mismatch while preserving the relative ordering within each model:

$$\hat{\phi}^{(m)}_j = \frac{\bar{\phi}^{(m)}_j}{\sum_{k=1}^{d} \bar{\phi}^{(m)}_k}. \tag{8}$$

Let $\hat{\mu}_j$ and $\hat{\sigma}_j$ be the mean and standard deviation of $\hat{\phi}^{(m)}_j$ across models. The coefficient of variation

$$\mathrm{CV}_j = \frac{\hat{\sigma}_j}{\hat{\mu}_j} \tag{9}$$

is invariant under rescaling of feature j's attributions, and the corrected stability index is the bounded transform

$$\mathrm{FSS}^*_j = \frac{1}{1+\mathrm{CV}_j} \in (0, 1]. \tag{10}$$

Perfect agreement gives $\mathrm{FSS}^*_j = 1$; the score decreases monotonically with relative disagreement and cannot become negative. The corrected composite S^*_j is formed exactly as in (5), with $\hat{\mu}_j$ replacing μ_j in the normalisation and FSS^*_j replacing FSS_j.

9) Individual-Patient Analysis

The entire construction applies unchanged at the level of a single patient by substituting the local absolute attributions

$|\phi_j(f^{(m)}, \mathbf{x})|$ for the cohort averages of (2). Cohort and patient rankings were compared by Spearman rank correlation, and each model's predicted probability for the patient was recorded to assess whether the models agree on the clinical decision as well as on its explanation.

4. Results

1) *Cohort Characteristics*

Fig. 1 summarises the cohort and Fig. 2 the correlation structure. Chest-pain type shows the strongest association with the outcome ($r = 0.43$), followed by maximum heart rate ($r = 0.42$), exercise-induced angina ($r = -0.44$), ST depression ($r = -0.43$), and the number of major vessels ($r = -0.39$). The relationship visible in Fig. 1(b) is pronounced: asymptomatic chest pain (type 0) dominates the negative class, while types 1 and 2 are concentrated among patients with disease.

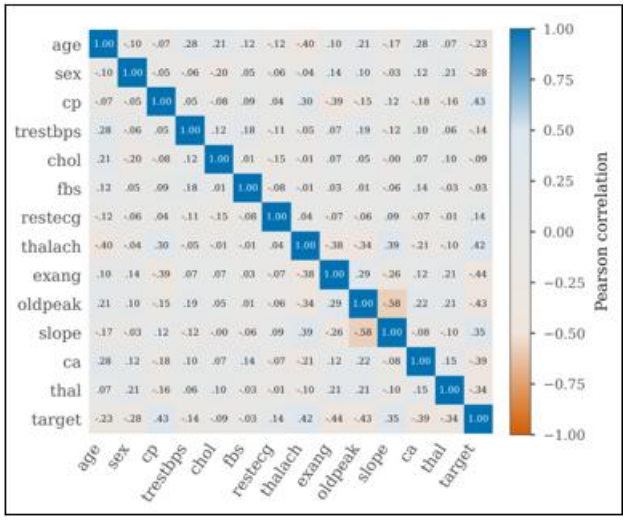

Figure 2: Pearson correlation matrix. Chest-pain type, maximum heart rate, exercise-induced angina, and ST depression show the strongest associations with the outcome; serum cholesterol is essentially uncorrelated with it.

Serum cholesterol, by contrast, is almost uncorrelated with the outcome ($r = -0.09$), which is worth noting in advance because it becomes prominent in one model's attribution profile in Section IV-C.

2) *Predictive Performance*

Table II reports held-out performance. Random Forest is the strongest model on both criteria, with accuracy 0.8361 and ROC-AUC 0.9188; XGBoost follows at 0.8197 and 0.8723, and Logistic Regression at 0.8033 and 0.8690. All three exceed the 54.5% no-skill floor by a wide margin.

Two features of the error profile deserve comment. First, all three models achieve identical recall of 0.68 on the negative class while differing on the positive class, so the ranking is driven entirely by how many diseased patients each model recovers: Random Forest identifies 32 of 33 (recall = 0.97), XGBoost 31 (0.94), and Logistic Regression 30 (0.91). Second, all three are markedly better at detecting disease than at excluding it, missing roughly a third of the healthy patients. In a screening context this asymmetry is the preferable direction, but it means the models function better as a rule-in than as a rule-out instrument.

The absolute differences are small in patient terms. The gap between the best and worst model is two reclassified patients out of 61, which on a test partition of this size is well within sampling variability. This is the empirical basis for the concern raised in Section II: selecting one of these models on accuracy grounds, and then treating its explanation as authoritative, rests on a distinction the data do not reliably support.

Table II: Held-Out Test Performance (n = 61: 28 Without Disease, 33 With Disease)

Model	Accuracy	No disease (class 0)			Disease (class 1)			Macro F_1	ROC-AUC
		Precision	Recall	F_1	Precision	Recall	F_1		
Logistic Regression	0.8033	0.86	0.68	0.76	0.77	0.91	0.83	0.8	0.869
Random Forest	0.8361	0.95	0.68	0.79	0.78	0.97	0.86	0.83	0.9188
XGBoost	0.8197	0.9	0.68	0.78	0.78	0.94	0.85	0.81	0.8723

3) *Cohort-Level Attribution*

Fig. 3 shows the five highest-ranked features for each model. All three place chest-pain type first, and both ensembles rank thalassaemia second and number of major vessels third. Beyond that the profiles diverge in ways that matter.

Logistic Regression ranks sex second, at 0.6625, while Random Forest places it eighth. XGBoost ranks chol fourth at 0.7030—higher than thalach, oldpeak, or exang—even though cholesterol has essentially no marginal association with the outcome in this cohort ($r = -0.09$). The most extreme divergence is age: Logistic Regression attributes it 0.0138, the smallest value in its entire profile, whereas XGBoost attributes it 0.4227, a factor of roughly thirty.

These are not small perturbations of a shared ranking; they are qualitative disagreements about which clinical variables matter. A study reporting only the Logistic Regression profile would conclude that sex is a dominant risk factor and age is irrelevant; a study reporting only the XGBoost profile would conclude close to the opposite. Both models would have been defensible selections on accuracy grounds.

The magnitudes across panels of Fig. 3 are not comparable, for the reason given in Section III-G: the Random Forest values are in probability units and the others in log-odds units. Only the within-panel ordering is meaningful.

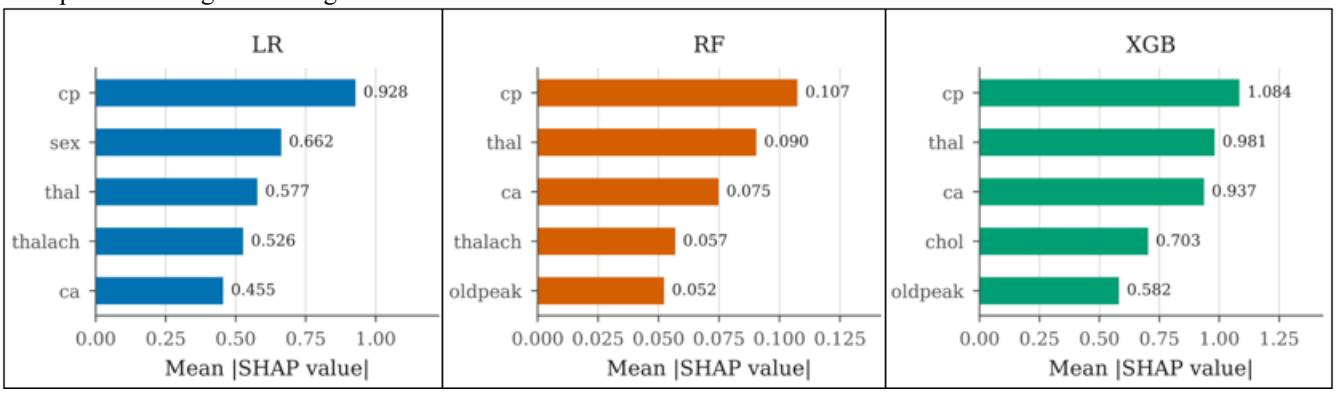

Figure 3: Top five features by mean absolute SHAP value for each model. All three rank chest-pain type first, but diverge sharply below that: sex is second for Logistic Regression and eighth for Random Forest, and chol is fourth for XGBoost despite being almost uncorrelated with the outcome. Magnitudes are not comparable across panels—Random Forest attributions are in probability units, the others in log-odds units.

4) Feature Stability: The Variance Artefact

Table III reports the full stability analysis under the variance formulation of Section III-F. Read as intended, it says that fbs is the most stable feature in the dataset (FSS = 0.9995) and cp the least (0.8166).

That reading is untenable. fbs is the weakest predictor in the study- its mean cross-model importance of 0.0327 is the lowest of the thirteen- and cp is the strongest and the only feature ranked first by all three models. A stability measure that ranks unanimous agreement about the leading predictor below near-indifference about the weakest one is not measuring agreement.

Fig. 4(a) shows the cause directly. Stability and importance lie almost exactly on a descending line: the Spearman rank correlation is $\rho = -0.973$ ($p = 2.6 \times 10^{-8}$) and the Pearson correlation is -0.955. This is the behaviour predicted by (7). The variation in relative disagreement across features is far too small to overcome the μ^2_j factor, so FSS is, to a very good approximation, a monotone decreasing transform of importance.

The consequence for the composite score is that the stability term contributes essentially nothing. The final ranking in Table III correlates with raw mean importance at $\rho = +0.995$ ($p = 3.9 \times 10^{-12}$), differing from a pure importance ordering only in the transposition of age and restecg. At the individual-patient level the collapse is complete: the composite ranking correlates with importance at $\rho = +1.000$, an exactly identical ordering. The stability analysis, as formulated, is an expensive way of re-deriving mean importance.

Table III: Cohort-Level Stability Under the Variance Formulation

Feature	LR	RF	XGB	Mean	σ^2	FSS
cp	0.9278	0.1074	1.0839	0.7064	0.1834	0.8166
thal	0.5772	0.0904	0.9809	0.5495	0.1325	0.8675
ca	0.455	0.0748	0.9367	0.4889	0.1244	0.8756
sex	0.6625	0.0287	0.5438	0.4116	0.0757	0.9243
chol	0.4415	0.0271	0.703	0.3906	0.0774	0.9226
thalach	0.5263	0.0569	0.46	0.3477	0.043	0.957
oldpeak	0.392	0.0523	0.5816	0.342	0.048	0.952
exang	0.4449	0.0513	0.4332	0.3098	0.0334	0.9666
slope	0.2706	0.0407	0.2827	0.198	0.0124	0.9876
age	0.0138	0.0227	0.4227	0.1531	0.0364	0.9636
restecg	0.2243	0.0145	0.2166	0.1518	0.0094	0.9906
trestbps	0.1557	0.0097	0.1844	0.1166	0.0059	0.9941
fbs	0.0553	0.0023	0.0405	0.0327	0.0005	0.9995

Rows are ordered by the composite score S_j. The most important feature (cp) receives the *lowest* stability score and the least important (fbs) the highest: Spearman ρ(FSS, Mean) = -0.973.

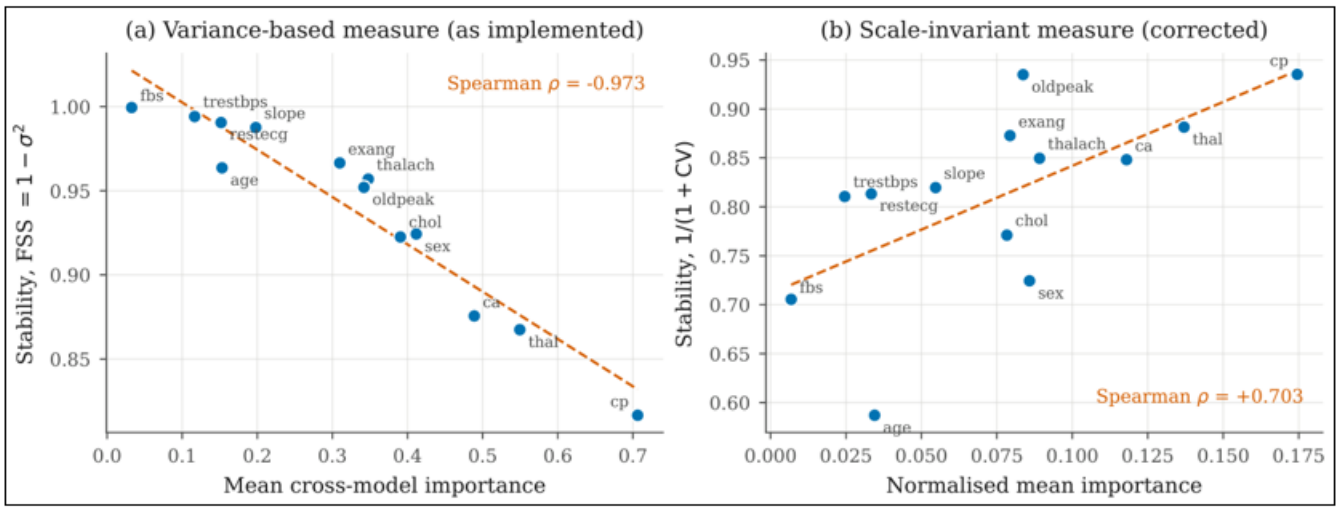

Figure 4: The stability measure before and after correction. (a) Under FSS = 1 − σ², stability is an almost perfect inverse function of importance (ρ = −0.973): the measure re-expresses importance instead of quantifying agreement. (b) Under the scale-invariant form 1/(1 + CV) applied to per-model normalised attributions, the two axes decouple (ρ = +0.703) and features become separable by agreement at comparable importance- oldpeak and ca are attributed consistently, whereas age, sex, and chol are strongly disputed

5) The Corrected Ranking

Table IV and Fig. 4(b) report the analysis repeated under the scale-invariant formulation of Section III-H. The correlation between stability and importance falls from ρ = −0.973 to ρ = +0.703, and the residual positive association is itself interpretable: features that all models consider important tend also to be those they agree about, which is the expected behaviour of a working agreement measure rather than an artefact of scale.

Stability now separates features that the variance formulation could not distinguish. Among predictors of comparable importance, oldpeak (CV = 0.069) and ca (0.179) are attributed consistently across all three models, whereas sex (0.380) and chol (0.297) are strongly disputed. The largest disagreement in the dataset is age, at CV = 0.703.

Fig. 5 shows the resulting movement. The top three positions- cp, thal, ca—are unchanged, which is a useful robustness result: these three features are both important and consistently attributed, and can be reported with confidence. Below them the ranking rearranges substantially. oldpeak rises from seventh to fourth and thalach from sixth to fifth, both promoted for consistency. sex falls from fourth to seventh and chol from fifth to eighth, both demoted because the models disagree about them. age falls from eleventh to twelfth, the lowest-ranked feature apart from fbs, reflecting the thirty-fold discrepancy between Logistic Regression and XGBoost.

The clinical reading is more defensible than the original. ST depression during exercise (oldpeak) is a well-established ischaemic marker and is promoted; serum cholesterol, which is almost uncorrelated with the outcome in this cohort and is emphasised only by the boosted model, is demoted. The corrected measure moves the ranking toward, rather than away from, established cardiological understanding.

Table IV: Cohort-Level Ranking Under the Corrected Formulation

Feature	$\hat{\mu_j}$	CV_j	FSS^*_j	S^*_j	Rank change
cp	0.1746	0.0691	0.9353	0.9741	=
thal	0.137	0.1345	0.8815	0.8183	=
ca	0.118	0.179	0.8482	0.7369	=
oldpeak	0.0837	**0.0694**	0.9351	0.649	3
thalach	0.0892	0.177	0.8496	0.6343	1
exang	0.0794	0.1457	0.8728	0.6085	2
sex	0.0858	0.3804	0.7244	0.5723	−3
chol	0.0783	0.2969	0.7711	0.5641	−3
slope	0.0547	0.22	0.8197	0.4991	=
restecg	0.0334	0.2296	0.8133	0.4203	=
trestbps	0.0246	0.2337	0.8105	0.3877	1
age	0.0345	**0.7034**	0.5871	0.3337	−1
fbs	0.0069	0.4174	0.7055	0.2822	=

Rank change is relative to the variance-based ranking of Table III. Spearman ρ(FSS*, μ̂) = +0.703, against −0.973 before correction.

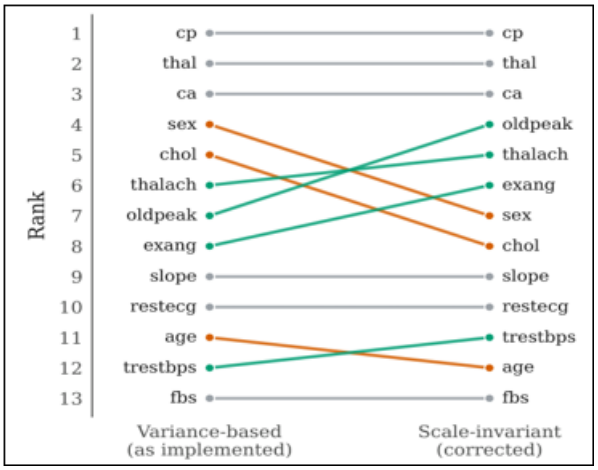

Figure 5: Movement in the cohort ranking after correcting the stability measure. The three leading features are unchanged. Features promoted for consistent attribution are shown

ascending; those demoted because the models disagree about them are shown descending.

6) *Individual-Patient Analysis*

The framework was applied to a representative record: a 60-year-old male with non-anginal chest pain (cp = 2), resting blood pressure 140 mm Hg, cholesterol 250 mg/dL, maximum heart rate 130, exercise-induced angina present, ST depression 2.3, upsloping ST segment, no major vessels affected, and a fixed-defect thalassaemia result.

a) The models disagree on the decision: Fig. 6(a) shows the outcome. Logistic Regression returns $P = 0.4987$, Random Forest 0.6224, and XGBoost 0.1995. Applying the standard 0.5 threshold, Random Forest predicts heart disease while the other two predict its absence- and Logistic Regression does so by a margin of 0.0013, so that a threshold shift of one part in a thousand would reverse it.

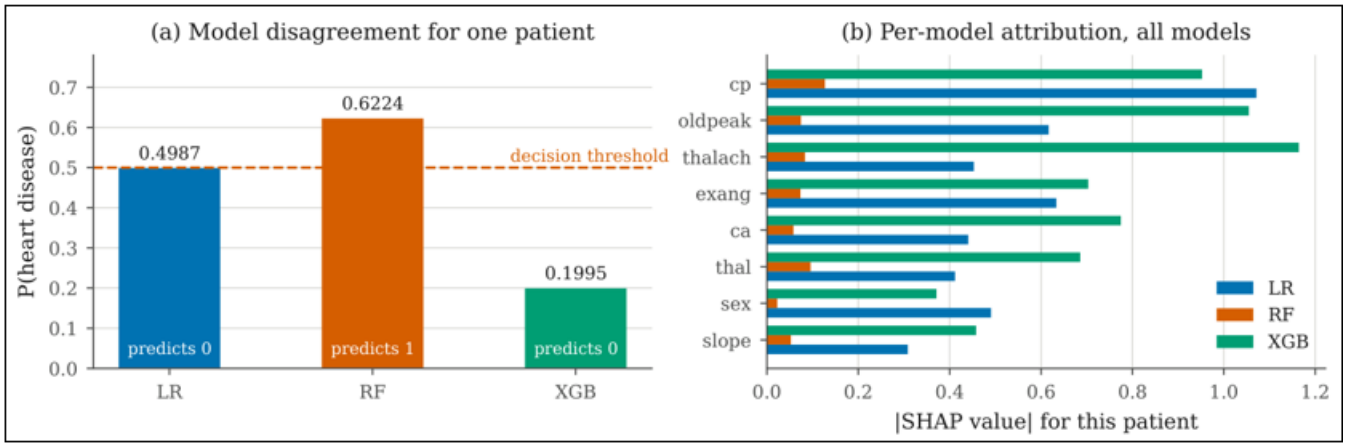

Figure 6: Individual-patient analysis. (a) The three models return probabilities spanning 0.20 to 0.62 and split on the resulting decision; Logistic Regression falls below the threshold by 0.0013. (b) Per-model absolute SHAP attributions for the eight features with the highest mean attribution, showing that the models disagree about the explanation as well as the prediction.

The spread is the more troubling quantity. Three models whose held-out accuracies differ by two reclassified patients assign this individual probabilities spanning 0.20 to 0.62, a range wide enough to encompass both confident exclusion and probable disease. Whatever number reaches the clinician is determined by a modelling choice that the aggregate metrics do not justify.

b) The explanations also disagree: Fig. 6(b) and Table V give the per-model attributions and the patient-level ranking. All three models place cp among the patient's leading factors, but they disagree beneath it: Logistic Regression attributes most weight to cp, exang, and oldpeak, whereas XGBoost leads with thalach (1.1642) and oldpeak (1.0548). For chol the values are 0.0204, 0.0056, and 0.5079: the boosted model treats this patient's cholesterol as a substantial factor while the other two treat it as almost irrelevant.

c) The patient ranking is not the cohort ranking: The patient-level composite ranking is cp, oldpeak, thalach, exang, ca, against the cohort ranking cp, thal, ca, sex, chol. Eleven of thirteen positions change and the rank correlation between the two is $\rho = 0.698$ (Fig. 7). oldpeak rises from seventh at cohort level to second for this patient—appropriately, since his ST depression of 2.3 is markedly abnormal—while sex and chol fall away.

This has a concrete design implication: an explanation presented to a clinician must be computed for the patient in front of them. Substituting the cohort ranking would have identified thal and ca as this patient's principal factors when in fact his exercise response, captured by oldpeak and thalach, is what the models react to.

Table V: Patient-Level Attribution and Composite Ranking

Feature	LR	RF	XGB	Mean	FSS	S_j
cp	1.0714	0.127	0.9523	0.7169	0.8237	0.9295
oldpeak	0.6167	0.0744	1.0548	0.5819	0.8392	0.8191
thalach	0.453	0.0833	1.1642	0.5668	0.7988	0.7898
exang	0.6334	0.0734	0.7032	0.47	0.9205	0.7548
ca	0.4406	0.0582	0.7747	0.4245	0.9143	0.713
thal	0.4118	0.0956	0.6857	0.3977	0.9419	0.7009
sex	0.4906	0.0228	0.3711	0.2948	0.9606	0.6195
slope	0.3084	0.0526	0.4578	0.273	0.972	0.6051
age	0.0088	0.0198	0.5553	0.1946	0.9349	0.5226
chol	0.0204	0.0056	0.5079	0.178	0.9455	0.5125
restecg	0.1955	0.0076	0.027	0.0767	0.9929	0.4439
trestbps	0.0809	0.01	0.0062	0.0324	0.9988	0.408
fbs	0.0348	0.0006	0.0325	0.0226	0.9998	0.3999

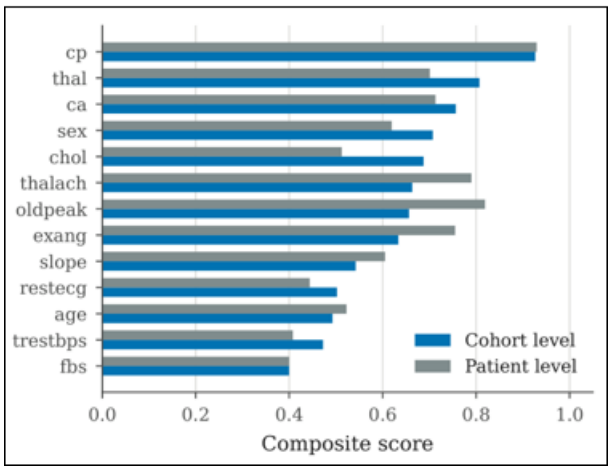

Figure 7: Composite score at cohort and patient level, ordered by the cohort ranking. Eleven of thirteen positions change between the two ($\rho = 0.698$): oldpeak and thalach rise sharply for this patient while sex and chol fall.

5. Discussion

1) ***A Stability Measure Must Be Scale-Invariant***

The central methodological finding is that the natural definition of cross-model stability is the wrong one. Cross-model variance appears to measure disagreement, but because variance scales with the square of the quantity being compared, it in fact measures magnitude. On these data the confound is almost total: $\rho = -0.973$ between stability and importance, and a composite score that reproduces the importance ordering at $\rho = +0.995$ at cohort level and exactly at patient level.

The failure mode is instructive because it is silent. The variance-based analysis produces a plausible-looking table, a defensible-looking ranking, and no diagnostic that anything is wrong; fbs scoring highest on stability reads as an unremarkable detail rather than as a symptom. Only the correlation between the two axes- a single number that the analysis does not by default compute- exposes it. We would suggest that any composite of importance and stability be accompanied by the rank correlation between its components, since a value near ±1 indicates that one component is redundant.

The correction is not merely cosmetic. Under the scale-invariant formulation the two axes decouple ($\rho = +0.703$), and the ranking changes in ways that are clinically meaningful: oldpeak, an established ischaemic marker, is promoted for consistent attribution, while chol and age, which the models dispute, are demoted. The corrected ranking is closer to cardiological understanding than the original, which is weak evidence in its favour but evidence nonetheless.

2) Model Disagreement Is the Finding, Not the Noise

The three models are close on aggregate metrics and far apart on nearly everything else. They disagree about which features matter- sex second for Logistic Regression and eighth for Random Forest; age varying thirty-fold between Logistic Regression and XGBoost- and about individual patients, returning probabilities from 0.20 to 0.62 for the same record and splitting on the decision.

It is tempting to treat this as noise to be resolved by picking a winner. That response is not available here, because the accuracy differences that would justify a choice amount to two patients out of sixty-one. The disagreement is better understood as a measurement of how much the available evidence actually determines, and reporting it is more honest than concealing it behind a selected model. The corrected stability index is precisely such a report: features with high FSS^* are conclusions the data support regardless of modelling choice, and features with low FSS^* are conclusions that belong to a model rather than to the data.

The chol case illustrates the practical value. Cholesterol is nearly uncorrelated with the outcome in this cohort ($r = -0.09$), yet XGBoost ranks it fourth. A single-model study selecting XGBoost would have reported serum cholesterol as a leading predictor of heart disease in this cohort—a claim the marginal statistics do not support and that likely reflects the boosted model's capacity to exploit a continuous variable with many candidate split points. The cross-model analysis flags it automatically through a high coefficient of variation.

3) Explanations Must Be Computed Per Patient

The cohort and patient rankings correlate at only $\rho = 0.698$, with eleven of thirteen positions changed, and the substitutions are substantive: this patient's risk is driven by his exercise response (oldpeak, thalach) rather than by the thal and ca findings that dominate at cohort level. Presenting him with the cohort explanation would have been actively misleading. Since TreeSHAP for a single record costs milliseconds, there is no practical reason for a decision-support interface to display anything else.

We note that the notebook implementation underlying this study initially reported identical cohort and patient rankings, an artefact of state carried over between computations; the correct patient-level values are those in Table V. The episode is a reminder that the divergence between global and local explanation is easy to lose accidentally, and worth checking explicitly.

4) Clinical Plausibility

The stable features accord with clinical understanding. Chest-pain type ranking first is expected, since anginal presentation is the primary symptomatic route to a coronary diagnosis. Thalassaemia result and number of major vessels- both derived from imaging and stress testing- are the most direct evidence in the dataset of underlying vascular pathology. ST depression and maximum heart rate capture the exercise response that stress testing is designed to elicit.

The disputed features are also intelligible. sex is a genuine risk factor, but as a binary variable its effect is easily absorbed into interactions by tree ensembles while remaining an explicit coefficient in a linear model, which plausibly explains why

Logistic Regression weights it far more heavily. age shows the reverse pattern: the linear model finds almost no independent linear effect once the exercise variables are present, whereas the boosted model exploits it in interactions. Neither is wrong; they are different decompositions of overlapping information, and the disagreement measure correctly identifies that the attribution is not determined by the data alone.

6. Limitations

Sample size and statistical uncertainty: With 303 records and a 61-record test partition, one reclassified patient shifts accuracy by 1.6 percentage points. All reported metrics derive from a single stratified split without confidence intervals, and the differences between the three models are not statistically distinguishable at this sample size. Repeated stratified cross-validation with interval estimates would be required to establish which differences are real; the arguments of this paper do not depend on the ordering, but any claim that one model is superior would.

Cohort scope: The Cleveland cohort is single-centre, historical, and predominantly male, and its 54.5% disease prevalence reflects a referred population rather than a screening one. Neither the absolute risk estimates nor the attribution profiles should be assumed to transfer to a general population, and no external validation cohort was available.

Encoding of categorical variables: cp, restecg, slope, ca, and thal are unordered categories encoded as consecutive integers. This imposes a spurious ordering on the linear model, which must fit a single monotone coefficient across categories, while tree models are unaffected. Some of the measured cross-model disagreement for these features is therefore attributable to the encoding rather than to genuine differences in what the models learned. One-hot encoding, with attributions summed within each variable, would remove this confound and is the most important refinement of the present analysis.

Three models is a small sample: The coefficient of variation is estimated from three values, so its sampling error is considerable and small differences in FSS^* should not be over-interpreted. A larger and more diverse model pool- support vector machines, *k*-nearest neighbours, regularised linear variants, and neural networks- would give a better-conditioned estimate, and bootstrap intervals on CV_j would make the uncertainty explicit.

Single patient: The individual-level analysis examines one synthetic record. It demonstrates that cohort and patient rankings can diverge sharply and that models can split on a decision, but it does not establish how often either occurs. Computing the distribution of cohort-to-patient rank correlation and of inter-model probability spread across the whole test set is a direct and inexpensive extension.

No calibration assessment: Discrimination was evaluated but calibration was not. Because Section IV-F interprets predicted probabilities as clinically meaningful quantities, this is a genuine gap: the Brier score and reliability diagrams [23], [24] would establish whether the probabilities are trustworthy in absolute terms, and the TRIPOD guideline [25] requires both.

Post-hoc explanation caveats: SHAP's standard estimators assume feature independence, which is violated by the correlated predictors in Fig. 2 [9]. More fundamentally, SHAP explains the model rather than the disease: a high attribution indicates reliance by the model, not a causal risk factor, and no therapeutic inference should be drawn from attribution magnitude [12].

7. Conclusion and Future Work

This paper presented an interpretable machine learning framework for heart disease prediction in which agreement between models is treated as a measurable quantity rather than assumed. Logistic Regression, Random Forest, and XGBoost were trained on the Cleveland dataset and explained with SHAP at both cohort and individual level, with Random Forest achieving the best discrimination (accuracy 0.8361, ROC-AUC 0.9188).

The principal methodological result is negative and then constructive. Defining cross-model stability as one minus the attribution variance produces a measure that is confounded with importance, because variance is not scale-invariant: the resulting score correlates with importance at $\rho = -0.973$, ranks the weakest predictor in the dataset as the most stable, and yields a composite ranking that reproduces raw importance at $\rho = +0.995$. The scale-invariant reformulation based on the coefficient of variation of per-model normalised attributions removes the confound ($\rho = +0.703$) and produces a ranking that separates consistently attributed predictors such as ST depression from disputed ones such as cholesterol and age.

The substantive result is that model choice matters more than the aggregate metrics suggest. Models separated by two reclassified patients disagree about feature importance by factors of up to thirty, assign the same individual probabilities from 0.20 to 0.62, and split on the resulting clinical decision. The cohort-level explanation does not transfer to the individual ($\rho = 0.698$). A single-model explanation of a heart disease classifier is therefore not sufficient evidence about the underlying predictors, and cross-model agreement should be reported alongside importance.

Future work follows from Section VI: one-hot encoding of the categorical predictors with attributions aggregated per variable; repeated stratified cross-validation with confidence intervals on both performance and CV_j; a larger and more diverse model pool; distributional analysis of cohort-to-patient divergence and inter-model spread across all test records; the addition of calibration assessment and post-hoc recalibration; and comparison of the proposed index against established stability statistics [18] and against the explanation-disagreement metrics of Krishna *et al.* [15].

Reproducibility

A fixed random seed (42) was used for partitioning and for the Random Forest fit. Models were implemented with scikit-learn [22] and XGBoost [21], and attributions computed with the shap library [5]. The Logistic Regression and Random Forest results, including every SHAP value reported here, reproduce exactly across platforms. XGBoost metrics are build-dependent, because row and column subsampling draw from a pseudo-random stream whose implementation has changed across releases and whose seed is not fixed in the reported configuration; the XGBoost figures given here are those of the original experiment, and a re-execution on a different platform may differ in the second decimal place. This does not affect the paper's conclusions, which concern the relationships between attribution profiles rather than the exact value of any single metric.

Acknowledgment

The authors thank the Department of Computer Science and Engineering, NITTTR Chennai, for its support of this work.

References

[1] G. A. Roth *et al.*, "Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study," *Journal of the American College of Cardiology*, vol. 76, no. 25, pp. 2982–3021, 2020.

[2] R. Detrano *et al.*, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *American Journal of Cardiology*, vol. 64, no. 5, pp. 304–310, 1989.

[3] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, "Heart Disease," *UCI Machine Learning Repository*, 1988.

[4] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 4765–4774.

[5] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020.

[6] L. S. Shapley, "A value for *n*-person games," in *Contributions to the Theory of Games II*, H. W. Kuhn and A. W. Tucker, Eds. Princeton, NJ, USA: Princeton University Press, 1953, pp. 307–317.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[8] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowledge and Information Systems*, vol. 41, no. 3, pp. 647–665, 2014.

[9] K. Aas, M. Jullum, and A. Løland, "Explaining individual predictions when features are dependent: more accurate approximations to Shapley values," *Artificial Intelligence*, vol. 298, art. 103502, 2021.

[10] M. Sundararajan and A. Najmi, "The many Shapley values for model explanation," in *Proc. 37th International Conference on Machine Learning (ICML)*, 2020, pp. 9269–9278.

[11] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods," in *Proc. AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2020, pp. 180–186.

[12] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[13] A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," in *Proc. AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 3681–3688.

[14] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," in *Proc. ICML Workshop on Human Interpretability in Machine Learning (WHI)*, 2018.

[15] S. Krishna *et al.*, "The disagreement problem in explainable machine learning: a practitioner's perspective," arXiv:2202.01602, 2022.

[16] M. Saarela and S. Jauhiainen, "Comparison of feature importance measures as explanations for classification models," *SN Applied Sciences*, vol. 3, art. 272, 2021.

[17] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: a study on high-dimensional spaces," *Knowledge and Information Systems*, vol. 12, no. 1, pp. 95–116, 2007.

[18] S. Nogueira, K. Sechidis, and G. Brown, "On the stability of feature selection algorithms," *Journal of Machine Learning Research*, vol. 18, no. 174, pp. 1–54, 2018.

[19] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed., 2022.

[20] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[21] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[22] F. Pedregosa *et al.*, "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[23] B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, and E. W. Steyerberg, "Calibration: the Achilles heel of predictive analytics," *BMC Medicine*, vol. 17, art. 230, 2019.

[24] E. W. Steyerberg *et al.*, "Assessing the performance of prediction models: a framework for traditional and novel measures," *Epidemiology*, vol. 21, no. 1, pp. 128–138, 2010.

[25] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement," *BMJ*, vol. 350, art. g7594, 2015.