

Research Paper



Equally accurate, differently explained: Stability of permutation feature importance across machine learning classifiers on clinical tabular data

Surjeet Dalal^{1*}, Mohd Akbar²

¹Professor & Head, Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Gurugram, Haryana, India, Visiting Professor, City Institute of Higher Learning (C-IHL), Canada.

²Senior Teaching Staff, Department of Computing and Information Sciences, School of Computing and Information Sciences, University of Technology and Applied Sciences, Muscat, Sultanate of Oman.

Article Info

Article History:

Received: 23 June 2026

Revised: 22 August 2026

Accepted: 13 September 2026

Published: 22 October 2026

Keywords:

Explainable Artificial Intelligence

Permutation Feature Importance

Explanation Stability

Rashomon Effect

Clinical Machine Learning



ABSTRACT

Feature-importance explanations are increasingly used to justify machine learning (ML) models in clinical decision support, yet they are usually reported from a single fitted model without evidence that they would recur if the model were refitted. This study quantifies the stability of permutation feature importance across five widely used classifiers, namely logistic regression (LR), radial-basis support vector machine (SVM), random forest (RF), histogram gradient boosting (GB) and a multilayer perceptron (MLP), on three public clinical tabular datasets: Wisconsin diagnostic breast cancer (WDBC), Pima diabetes and Cleveland heart disease. Using 10×5 -fold stratified cross-validation (50 fits per model and dataset), we measured discrimination and calibration and computed AUC-based permutation importance on every held-out fold. Stability was summarised with pairwise Spearman correlation, top-5 Jaccard overlap and top-feature recurrence. On WDBC, LR, SVM, RF and GB were statistically indistinguishable in AUC (0.990–0.995; corrected resampled t-test $p \geq 0.089$), yet their explanation stability ranged from 0.32 (RF) to 0.61 (SVM), these differences were significant (Holm-adjusted Wilcoxon $p < 0.001$), and the models disagreed on the most important feature. Stability was highest where predictors were few and weakly correlated (Pima, 0.70–0.87, with glucose ranked first in every fit of four models) and lowest where many near-redundant features were present. Reducing the training data to 25% lowered AUC by at most 0.03 for the four well-behaved models but reduced stability markedly (e.g., 0.90 to 0.68 for LR on Pima), whereas degenerate MLP fits produced spuriously stable importance. Predictive accuracy therefore does not certify explanation reliability, and importance-based explanations should be reported together with resampling-based stability estimates.

Corresponding Author:

Surjeet Dalal

Professor & Head, Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Gurugram, Haryana, India, Visiting Professor, City Institute of Higher Learning (C-IHL), Canada.

Email: sdalal@ggn.amity.edu

Copyright © 2026 The Author(s). This is an open access article distributed under the Creative Commons Attribution License, (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

Machine learning (ML) models trained on routinely collected clinical variables now match or exceed conventional risk scores on many diagnostic and prognostic tasks, and explanation methods have become the principal means by which such models are justified to clinicians, regulators and patients [1], [2]. Among these methods, feature-importance measures are the most widely reported: they summarise which inputs a model relies on and are frequently read as statements about the underlying clinical problem. Permutation feature importance, introduced for random forests by Breiman [3] and generalised as model reliance by Fisher et al. [4], is particularly popular because it is model-agnostic, is expressed in the units of a performance metric, and is available in standard software [5].

An explanation, however, is only as trustworthy as it is reproducible. Feature importance is usually computed once, on one fitted model, and presented as a property of the problem. Yet a fitted model is itself a random quantity that depends on the particular training sample, and many different models can reach essentially the same accuracy while relying on different variables, a phenomenon known as the Rashomon effect [4], [6], [7]. Related concerns have been raised under the headings of predictive multiplicity [8], underspecification [9] and disagreement between explanation methods [10]. In health care, Ghassemi et al. [11] argued that current explanation practice offers a false sense of assurance, and Rudin [12] recommended inherently interpretable models for high-stakes decisions. Methodological pitfalls of permutation-based importance are also well documented: correlated predictors distort importance and force extrapolation [13], [14], [15], [16].

What remains insufficiently quantified is how much the importance ranking of a given model class changes under the ordinary sampling variability of a standard validation protocol, and whether this explanation stability is predictable from the accuracy figures that authors already report. If two classifiers are statistically indistinguishable in discrimination, is it safe to assume that their explanations are equally reliable? The feature-selection literature has long treated stability as a quality criterion in its own right [17], [18], but routine reporting of ML explanations rarely includes it, and current clinical reporting guidance emphasises discrimination and calibration [19], [20].

This paper addresses that gap with a controlled, fully reproducible benchmark. Five classifiers representing linear, kernel, bagged-tree, boosted-tree and neural model families were evaluated on three public clinical datasets of contrasting dimensionality and feature redundancy. Within a repeated cross-validation design, permutation importance was computed on every held-out fold, so that each model-dataset combination yielded 50 independent-fit explanations. The main contributions are as follows:

A quantitative comparison of predictive performance, calibration and explanation stability for five classifier families under identical resampling, showing that models with statistically indistinguishable AUC can differ significantly in the stability of their explanations.

Evidence that explanation stability depends strongly on feature redundancy and training-set size, and that it saturates far more slowly than discrimination as data are added.

Identification of a failure mode in which degenerate models yield near-constant, and therefore spuriously stable, importance vectors, implying that stability must be interpreted together with performance.

Practical reporting recommendations for importance-based explanations in clinical ML, grounded in the observed results.

The remainder of the paper is organised as follows. Section 2 reviews related work. Section 3 describes the datasets, models, explanation procedure, stability metrics and statistical analysis. Section 4 presents the results, Section 5 discusses their implications and limitations, and Section 6 concludes the study.

2. RELATED WORK

2.1 Permutation Feature Importance and Its Pitfalls

Permutation importance measures the loss in predictive performance when the values of one feature are randomly shuffled, breaking its association with the outcome while keeping its marginal distribution [3]. Fisher et al. [4] formalised this quantity as model reliance and extended it to model class reliance, the range of reliance values attainable by all well-performing models in a class. Strobl et al. [13] showed that impurity-based importances in random forests are biased towards variables with many categories, and later proposed conditional permutation schemes for correlated predictors [14]. Hooker et al. [15] demonstrated that unrestricted permutation forces models to extrapolate into regions without data when features are dependent, which can make importance values reflect model behaviour off the data manifold rather than genuine reliance. Molnar et al. [16] catalogued these and related pitfalls, including the effect of feature dependence and the need to quantify uncertainty in interpretation methods.

2.2 Robustness and Stability of Explanations

Alvarez-Melis and Jaakkola [21] showed that popular local explanation methods, such as LIME [22] and SHAP [23], can change substantially under small input perturbations. Krishna et al. [10] documented frequent disagreement between explanation methods applied to the same model and reported that practitioners resolve such conflicts in ad hoc ways. In the neighbouring field of feature selection, stability has long been recognised as a criterion distinct from accuracy: Kalousis et al. [17] measured the sensitivity of selected feature subsets to training-set perturbations, and Nogueira et al. [18] established axiomatic requirements for stability measures. These studies, however, concern selection procedures rather than the importance explanations produced by fitted predictive models.

2.3 Multiplicity of Equally Accurate Models

Breiman [6] described the Rashomon effect, whereby many models of similar accuracy tell different stories about the data. Marx et al. [8] formalised predictive multiplicity, D'Amour et al. [9] showed that pipelines with equivalent held-out performance can behave very differently in deployment, and Semenova et al. [7] linked large Rashomon sets to the existence of simpler models. Doshi-Velez and Kim [24] argued that interpretability claims require rigorous evaluation. Together, these works imply that explanation reliability cannot be inferred from accuracy alone, but empirical evidence quantifying the gap for standard classifiers on clinical tabular data remains limited.

2.4 Research Gap

Existing benchmarks typically compare classifiers on discrimination, occasionally on calibration [25], [26], and seldom on the reproducibility of their explanations. Studies that do consider explanation robustness usually examine a single model or input-level perturbations rather than refitting under a standard validation protocol. The present study combines a rigorous resampling design, corrected statistical inference [27], [28] and several complementary stability measures to compare explanation stability across model families, datasets and training-set sizes.

3. MEATHODOLOGY

The study pipeline is summarised in Figure 1. All preprocessing was fitted within training folds only, so that no information from test folds leaked into model fitting or imputation.

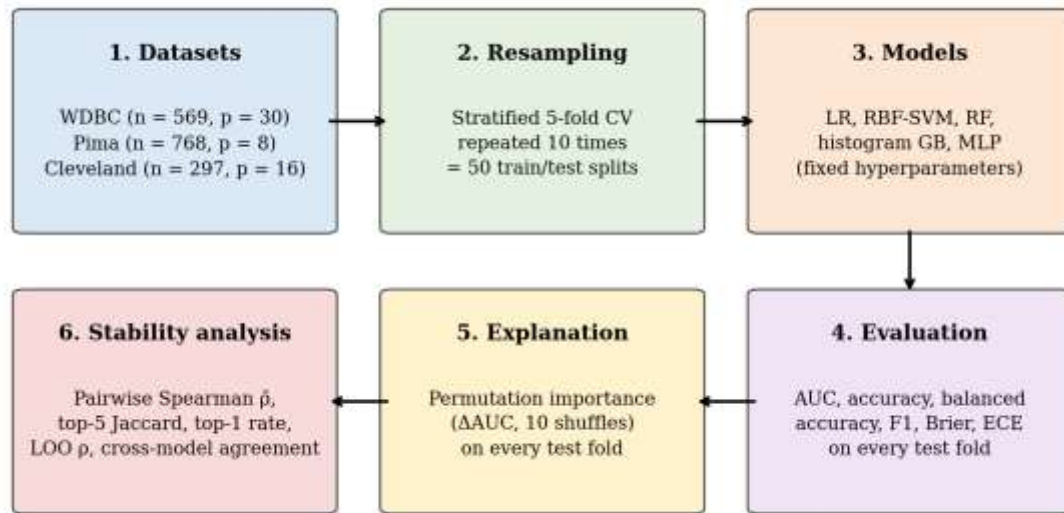


Figure 1. Overview of the Experimental Pipeline. Every Model is refitted on each of the 50 Training Folds, and Performance and Permutation Importance Are Computed on the Corresponding Held-Out Fold

3.1 Datasets

Three public clinical datasets with binary outcomes were selected to span different sample sizes, dimensionalities and degrees of feature redundancy Table 1. The Wisconsin diagnostic breast cancer (WDBC) dataset [29] contains 30 features describing cell nuclei in digitised fine-needle aspirate images, formed from the mean, standard error and worst value of ten nuclear characteristics; the positive class is malignancy. The Pima Indians diabetes dataset [30] describes 768 women aged 21 years or older with eight clinical measurements; physiologically impossible zero values of glucose, blood pressure, skin thickness, insulin and body mass index (652 values in total) were treated as missing. The Cleveland heart disease dataset [31] was used in its widely distributed 303-record form with 13 predictors; six records with missing values in the number of fluoroscopy-coloured vessels (Ca) or the thallium test result (Thal) were removed, and the categorical variables chest-pain type and Thal were one-hot encoded with the first level dropped, giving 16 features. Feature redundancy was quantified by the mean absolute pairwise Spearman correlation and the number of feature pairs with $|\rho| > 0.9$.

Table 1. Characteristics of the Datasets Used in the Study

Dataset	Samples (n)	Features (p)	Positive Class, n (%)	Missing Values	Mean $ \rho $ Between Features	Pairs with $ \rho > 0.9$
WDBC [29]	569	30	Malignant, 212 (37.3)	None	0.42	26 / 435
Pima [30]	768	8	Diabetes, 268 (34.9)	652 cells (zero-coded)	0.22	0 / 28
Cleveland [31]	297	16*	Heart disease, 137 (46.1)	6 records removed	0.15	0 / 120

*13 original predictors; 16 after one-hot encoding of chest-pain type and thallium test result.

3.2 Models and Preprocessing

Five classifiers representing distinct inductive biases were implemented in scikit-learn [5]: L2-regularised logistic regression (LR); a support vector machine with radial-basis kernel (SVM) [32], whose scores were converted to probabilities by Platt scaling; a random forest (RF) [3]; histogram-based gradient boosting (GB) [33], [34]; and a multilayer perceptron (MLP) [35]. Hyperparameters were fixed a priori at widely used values Table 2 rather than tuned, so that differences in explanation stability could not be produced by fold-specific tuning; nested tuning inside each of the 50 folds would add a further source of variability and is discussed as a limitation. Every model was wrapped in a pipeline that applied median imputation and, for LR, SVM and MLP, standardisation, both fitted on the training fold only.

Table 2. Model Configurations (Fixed Across All Datasets and Folds; Random Seed = 42)

Model	Family	Configuration	Preprocessing
LR	Linear	L2 penalty, $C = 1.0$, L-BFGS solver, max. 5000 iterations	Median imputation, standardisation
SVM	Kernel	RBF kernel, $C = 1.0$, $\gamma =$ "scale", Platt-scaled probabilities	Median imputation, standardisation
RF	Bagged trees	300 trees, $\sqrt{p}$ features per split, unrestricted depth	Median imputation
GB	Boosted trees	Histogram GB, 200 iterations, learning rate 0.05, 31 leaves	Median imputation
MLP	Neural network	Hidden layers (32, 16), ReLU, Adam, $\alpha = 10^{-3}$, early stopping on 10% validation split	Median imputation, standardisation

3.3 Evaluation Protocol and Performance Metrics

Each model was evaluated with stratified 5-fold cross-validation repeated 10 times with different shuffles, giving $T = 50$ train-test splits per dataset; the same splits were used for all models so that comparisons are paired. On each test fold we recorded the area under the receiver operating characteristic curve (AUC) [36], accuracy, balanced accuracy and F1-score at a 0.5 threshold, the Brier score [37], and the expected calibration error (ECE) [38] computed with ten equal-width bins: [1] where n_b is the number of test samples whose predicted probability falls in bin b , and $\bar{y}_b$ and $\bar{p}_b$ are the observed event rate and mean predicted probability in that bin.

3.4 Permutation Feature Importance

For each fitted model $f(t)$ and its held-out fold $(X(t), y(t))$, the importance of feature j was computed as the mean decrease in test AUC over $K = 10$ random permutations of column j : [2] where $X_{\pi k}$ denotes the test matrix with column j permuted. Computing importance on held-out data, rather than on the training data, measures reliance that generalises and avoids rewarding overfitting [4], [16]. Negative values, which arise when permuting a feature slightly improves test AUC by chance, were retained because they indicate the noise floor of the estimate. Each model-dataset combination thus produced an importance matrix of size $50 \times p$.

3.5 Stability Metrics

Let $I(s)$ and $I(t)$ be the importance vectors from two fits of the same model on the same dataset. Four complementary measures were computed. First, rank stability was defined as the mean Spearman correlation over all $T(T-1)/2 = 1225$ pairs of fits: [3] Second, top-k stability was measured by the mean pairwise Jaccard index of the sets S_k of the $k = 5$ highest-ranked features, which focuses on the head of the ranking that is usually reported: [4] Third, the top-1 recurrence rate was the proportion of fits in which the single most important feature equalled the modal top feature across all 50 fits. Fourth, to obtain one stability value per fit for paired statistical testing, the leave-one-out (LOO) stability of fit t was defined as the Spearman correlation between $I(t)$ and the mean importance vector of the remaining 49 fits. The share of negative importance values was additionally reported as an indicator of estimation noise. Cross-model agreement was quantified by the Spearman correlation between the mean importance vectors of two models on the same dataset.

3.6 Statistical Analysis

Differences among the five models were first assessed with the Friedman test [39], treating the 50 folds as blocks, separately for AUC and for LOO stability. Because resampled folds share training data, tests that treat folds as independent are anti-conservative [27], [28]; AUC differences between the best model and each competitor were therefore also tested with the Nadeau–Bengio corrected resampled t-test: [5] where $\bar{d}$ and $\hat{\sigma}^2$ are the mean and variance of the per-fold AUC differences, $T = 50$ and $n_{te}/n_{tr} = 1/4$ for 5-fold cross-validation, with $T - 1$ degrees of freedom. Pairwise differences in LOO stability were tested with the Wilcoxon signed-rank test [40] with Holm correction over the ten model pairs [41]. As LOO values of different fits are not strictly independent, these p-values are interpreted as descriptive evidence rather than exact error rates. The significance level was 0.05.

3.7 Training-Set Size Experiment and Implementation

To examine how explanation stability depends on the amount of training data, a second experiment used 5×5 -fold stratified cross-validation (25 fits) in which each training fold was stratified-subsampled to 25%, 50% or 100% of its size while the test fold was kept intact. All experiments were run in Python 3 with scikit-learn 1.8.0 [5], SciPy 1.17.1 and pandas 3.0.2 on a single CPU core with a fixed random seed (42). The datasets are public and the complete code is available from the corresponding author, so that every number reported below can be regenerated.

4. RESULTS AND DISCUSSION

4.1 Predictive Performance and Calibration

Table 3 reports discrimination and calibration averaged over the 50 test folds. On WDBC, LR and SVM reached the highest AUC (0.995), closely followed by GB (0.992) and RF (0.990). On Pima and Cleveland, LR achieved the highest AUC (0.835 and 0.901), with SVM and RF within 0.007. LR and SVM were also among the best calibrated models, whereas the MLP had the highest Brier score and ECE on WDBC and Cleveland. The MLP failed on Pima: 29 of its 50 fits predicted no positive cases at the 0.5 threshold ($F1 = 0$), and its AUC was low and highly variable (0.656 ± 0.160). This behaviour is attributable to early stopping on a small validation split and is retained in the analysis as an instructive failure case rather than tuned away post hoc.

Table 3. Predictive Performance and Calibration (Mean $\pm$ SD over 50 Test Folds of 10×5 -Fold Cross-Validation). Lower Brier Score and ECE Indicate Better Calibration

Dataset	Model	AUC	Accuracy	Balanced Accuracy	F1-Score	Brier Score	ECE
WDBC	LR	0.995 ± 0.005	0.979 ± 0.014	0.975 ± 0.018	0.971 ± 0.020	0.020 ± 0.008	0.032 ± 0.007

Dataset	Model	AUC	Accuracy	Balanced Accuracy	F1-Score	Brier Score	ECE
	SVM	0.995 ± 0.005	0.972 ± 0.014	0.970 ± 0.016	0.962 ± 0.019	0.021 ± 0.008	0.036 ± 0.010
	RF	0.990 ± 0.008	0.959 ± 0.020	0.954 ± 0.024	0.944 ± 0.028	0.032 ± 0.009	0.050 ± 0.012
	GB	0.992 ± 0.006	0.967 ± 0.015	0.962 ± 0.018	0.956 ± 0.021	0.028 ± 0.012	0.033 ± 0.011
	MLP	0.981 ± 0.012	0.926 ± 0.026	0.904 ± 0.035	0.890 ± 0.043	0.061 ± 0.016	0.102 ± 0.024
Pima	LR	0.835 ± 0.028	0.767 ± 0.029	0.718 ± 0.036	0.624 ± 0.055	0.157 ± 0.014	0.081 ± 0.021
	SVM	0.829 ± 0.025	0.758 ± 0.026	0.708 ± 0.031	0.609 ± 0.049	0.163 ± 0.012	0.090 ± 0.023
	RF	0.829 ± 0.027	0.763 ± 0.025	0.727 ± 0.028	0.641 ± 0.040	0.159 ± 0.012	0.072 ± 0.019
	GB	0.810 ± 0.027	0.750 ± 0.028	0.720 ± 0.033	0.633 ± 0.046	0.186 ± 0.018	0.142 ± 0.021
	MLP	0.656 ± 0.160	0.692 ± 0.054	0.577 ± 0.099	0.231 ± 0.287	0.215 ± 0.045	0.128 ± 0.036
Cleveland	LR	0.901 ± 0.044	0.835 ± 0.054	0.831 ± 0.056	0.811 ± 0.068	0.123 ± 0.032	0.112 ± 0.028
	SVM	0.895 ± 0.049	0.828 ± 0.053	0.824 ± 0.055	0.806 ± 0.066	0.129 ± 0.032	0.119 ± 0.028
	RF	0.895 ± 0.045	0.800 ± 0.055	0.799 ± 0.056	0.781 ± 0.062	0.133 ± 0.026	0.123 ± 0.029
	GB	0.878 ± 0.044	0.800 ± 0.051	0.797 ± 0.051	0.779 ± 0.057	0.148 ± 0.036	0.140 ± 0.039
	MLP	0.862 ± 0.063	0.756 ± 0.066	0.764 ± 0.061	0.768 ± 0.047	0.175 ± 0.027	0.170 ± 0.034

The Friedman test rejected equal AUC across the five models on all three datasets [Table 4](#). However, the corrected resampled t-test, which accounts for the overlap between training folds, showed that the leading models were not distinguishable: on WDBC, SVM did not differ significantly from LR ($\Delta\text{AUC} = 0.0003$, $p = 0.816$), GB (0.0031 , $p = 0.089$) or RF (0.0053 , $p = 0.183$); on Pima and Cleveland, LR did not differ from SVM or RF (all $p \geq 0.418$). Only the MLP on WDBC and Pima, and GB on Pima ($p = 0.049$), were significantly worse than the best model. For practical purposes, therefore, LR, SVM, RF and GB form a set of equally accurate models on WDBC and Cleveland.

Table 4. Statistical Comparison of AUC: Friedman Test across Five Models and Nadeau–Bengio Corrected Resampled T-Test between the Best Model and Each Competitor (T = 50 Folds)

Dataset	Friedman χ^2 (df = 4)	P-Value	Best Model	Comparison	ΔAUC	Corrected p
WDBC	104.34	1.2×10^{-21}	SVM	vs LR	0.0003	0.816

Dataset	Friedman χ^2 (df = 4)	P-Value	Best Model	Comparison	Δ AUC	Corrected p
				vs RF	0.0053	0.183
				vs GB	0.0031	0.089
				vs MLP	0.0139	0.015*
Pima	79.12	2.7×10^{-16}	LR	vs SVM	0.0067	0.418
				vs RF	0.0060	0.499
				vs GB	0.0250	0.049*
				vs MLP	0.1799	0.031*
Cleveland	42.25	1.5×10^{-8}	LR	vs SVM	0.0061	0.498
				vs RF	0.0060	0.584
				vs GB	0.0229	0.131
				vs MLP	0.0390	0.187

*Significant at $\alpha = 0.05$. The Friedman test treats folds as independent blocks and is therefore anti-conservative for resampled data [27].

4.2 Within-Model Explanation Stability

Explanation stability differed markedly between datasets and between models Table 5. Pima produced the most stable explanations ($\bar{\rho} = 0.70$ – 0.87), and glucose was the most important feature in every one of the 50 fits for LR, SVM, RF and GB. Cleveland showed intermediate stability ($\bar{\rho} = 0.49$ – 0.57), with the number of fluoroscopy-coloured vessels (Ca) ranked first in 94–96% of the fits of the four non-neural models. WDBC, despite yielding the highest accuracies of the study, produced the least stable explanations ($\bar{\rho} = 0.32$ – 0.61), and no feature was ranked first in more than 70% of the fits of any model.

Table 5. Explanation Stability of Permutation Importance over 50 Fits (Mean $\pm$ SD). Higher Values Indicate More Stable Explanations, Except for the Negative Importance Share, Which Indicates Estimation Noise

Dataset	Model	$\bar{\rho}$ (Pairwise Spearman)	Top-5 Jaccard	Top-1 Recurrence	L00 Stability	Negative Importance Share
WDBC	LR	0.57 ± 0.21	0.34 ± 0.18	0.46	0.73 ± 0.15	0.18
	SVM	0.61 ± 0.18	0.34 ± 0.18	0.70	0.76 ± 0.11	0.16
	RF	0.32 ± 0.36	0.38 ± 0.25	0.52	0.50 ± 0.32	0.32
	GB	0.48 ± 0.16	0.48 ± 0.18	0.40	0.62 ± 0.11	0.25
	MLP	0.43 ± 0.22	0.34 ± 0.21	0.30	0.63 ± 0.14	0.33
Pima	LR	0.87 ± 0.11	0.84 ± 0.17	1.00	0.92 ± 0.07	0.22
	SVM	0.86 ± 0.09	0.87 ± 0.17	1.00	0.91 ± 0.05	0.15
	RF	0.82 ± 0.11	0.73 ± 0.17	1.00	0.89 ± 0.07	0.18
	GB	0.76 ± 0.14	0.66 ± 0.17	1.00	0.81 ± 0.12	0.24
	MLP	0.70 ± 0.22	0.72 ± 0.19	0.98	0.81 ± 0.19	0.30
Cleveland	LR	0.55 ± 0.21	0.52 ± 0.19	0.94	0.70 ± 0.17	0.22
	SVM	0.50 ± 0.21	0.41 ± 0.18	0.96	0.68 ± 0.15	0.25

Dataset	Model	$\bar{P}$ (Pairwise Spearman)	Top-5 Jaccard	Top-1 Recurrence	LOO Stability	Negative Importance Share
	RF	0.57 ± 0.18	0.48 ± 0.17	0.96	0.72 ± 0.13	0.25
	GB	0.49 ± 0.22	0.43 ± 0.16	0.94	0.65 ± 0.19	0.20
	MLP	0.50 ± 0.18	0.44 ± 0.17	0.66	0.70 ± 0.12	0.26

The relationship between accuracy and stability is shown in Figure 2. Across the five models, mean AUC and mean stability were positively rank-correlated on WDBC and Pima (Spearman $\rho = 0.90$ on both) and moderately on Cleveland ($\rho = 0.60$), largely because the weakest model, the MLP, was also among the least stable. Within the set of equally accurate models, however, stability varied widely. On WDBC, the AUCs of LR, SVM, RF and GB lie within 0.005 of one another, whereas their $\bar{p}$ values range from 0.32 to 0.61, almost a two-fold difference.

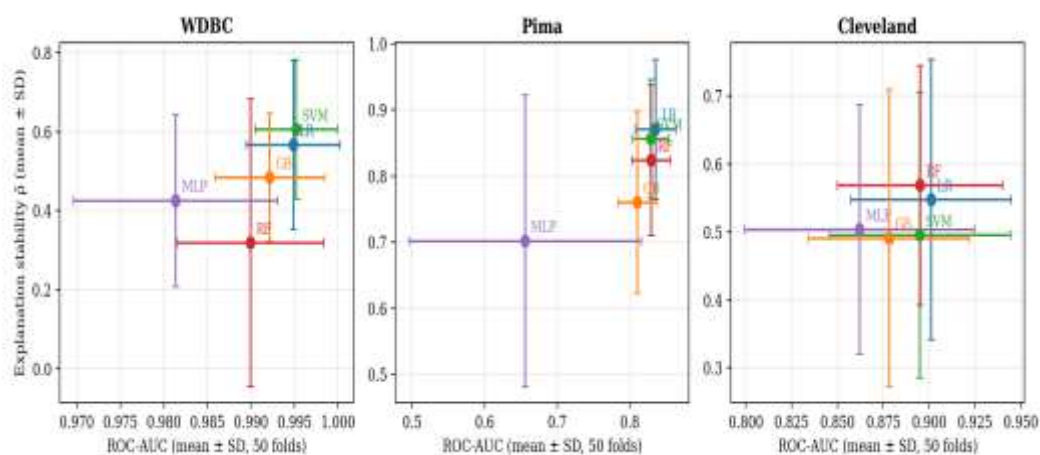


Figure 2. Discrimination (X-Axis, AUC) Versus Explanation Stability (Y-Axis, Mean Pairwise Spearman Correlation $\bar{p}$) for Each Model and Dataset. Error Bars Show One Standard Deviation across the 50 Folds (x) and Across the 1225 Pairs of Fits (y)

The two rank-based metrics also disagreed in an informative way. RF on WDBC had the lowest $\bar{p}$ (0.32) but a top-5 Jaccard index (0.38) comparable to or higher than that of LR and SVM (0.34). Its importance was concentrated on a few features, while the remaining features had near-zero importance whose ordering was dominated by noise, as reflected in the highest share of negative importance values (0.32). Whole-ranking and top-k stability therefore capture different aspects of reliability and should both be reported.

4.3 Statistical Comparison of Stability

The Friedman test on LOO stability rejected equal stability on WDBC ($\chi^2 = 45.20$, $p = 3.6 \times 10^{-9}$) and Pima ($\chi^2 = 34.12$, $p = 7.1 \times 10^{-7}$) but not on Cleveland ($\chi^2 = 6.63$, $p = 0.157$). The significant Holm-adjusted pairwise differences are listed in

Table 6. On WDBC, SVM and LR were significantly more stable than RF, GB and the MLP, with the largest difference between SVM and RF ($\Delta\text{LOO} = 0.26$, $p < 0.001$). This is the central result of the study: SVM and RF were indistinguishable in AUC ($p = 0.183$) but differed significantly and substantially in how reproducibly they ranked the features.

Table 6. Friedman Test on Leave-One-Out (LOO) Stability and Pairwise Differences That Remained Significant After Holm Correction (Wilcoxon Signed-Rank Test, Ten Pairs per Dataset)

Dataset	Friedman χ^2 (df = 4)	P-Value	Mean Rank (1 = Most Stable)	Significant Pairs (Δ LOO, Adjusted p)
WDBC	45.20	3.6×10^{-9}	SVM 2.04, LR 2.36, MLP 3.32, GB 3.58, RF 3.70	SVM > RF (0.26, < 0.001); SVM > GB (0.14, < 0.001); SVM > MLP (0.13, < 0.001); LR > RF (0.22, < 0.001); LR > GB (0.10, 0.002); LR > MLP (0.10, 0.017)
Pima	34.12	7.1×10^{-7}	LR 2.32, SVM 2.65, RF 2.79, MLP 3.28, GB 3.96	LR > GB (0.11, < 0.001); SVM > GB (0.10, < 0.001); RF > GB (0.08, < 0.001); LR > MLP (0.11, 0.002); SVM > MLP (0.10, 0.026)
Cleveland	6.63	0.157	RF 2.68, LR 2.78, MLP 3.00, SVM 3.14, GB 3.40	RF > GB (0.07, 0.032)

4.4 Agreement between Models

Figure 3 shows the Spearman correlation between the mean importance vectors of each pair of models. Agreement was high among the tree ensembles and SVM on Pima (0.93–0.98) and between RF and GB on Cleveland (0.91), but only moderate on WDBC, where the four equally accurate models agreed at $\rho = 0.49$ –0.85. The MLP was almost uncorrelated with all other models on WDBC (–0.05 to 0.19) and on Pima (0.05 to 0.64), consistent with its weaker and, on Pima, degenerate fits.

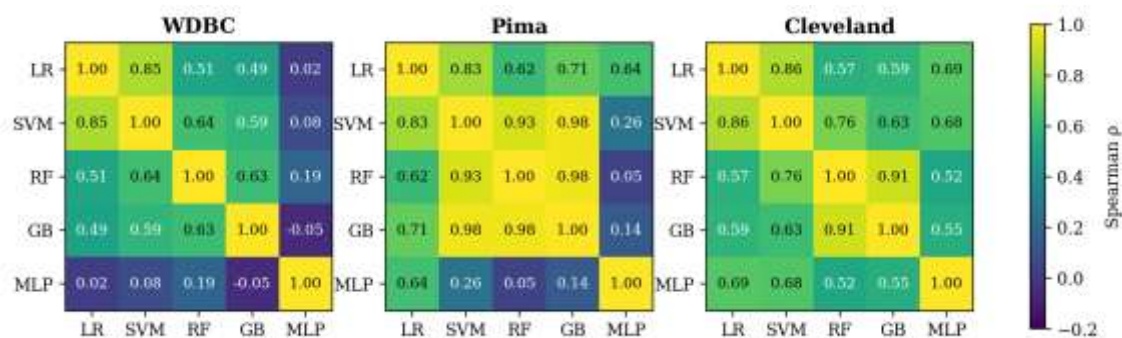


Figure 3. Cross-Model Agreement: Spearman Correlation Between the Mean Permutation-Importance Vectors (Averaged over 50 Fits) of Each Pair of Models

Table 7 lists the three highest-ranked features of each model. On Pima, all models agreed on glucose as the dominant predictor, followed by body mass index, in line with established clinical knowledge. On Cleveland, all five models ranked Ca first, but they differed in the secondary predictors: the linear and kernel models emphasised non-anginal chest pain and a reversible thallium defect, whereas the tree ensembles emphasised a normal thallium result, ST depression (Oldpeak) and sex. On WDBC, the models did not even agree on the leading feature: LR, SVM and MLP ranked worst texture first, RF ranked worst area first and GB ranked worst concave points first.

Table 7. The Three Most Important Features (by Mean Permutation Importance over 50 Fits) for Each Model and Dataset

Dataset	LR	SVM	RF	GB	MLP
WDBC	worst texture; radius error; worst symmetry	worst texture; worst radius; worst smoothness	worst area; worst concave points; worst perimeter	worst concave points; worst perimeter; worst area	worst texture; mean perimeter; mean radius
Pima	Glucose; BMI; Pregnancies	Glucose; BMI; Age	Glucose; BMI; Age	Glucose; BMI; Age	Glucose; Pregnancies; Pedigree
Cleveland	Ca; ChestPain (non-anginal); Thal (reversible)	Ca; ChestPain (non-anginal); Thal (reversible)	Ca; Thal (normal); Oldpeak	Ca; Sex; Thal (normal)	Ca; Sex; Thal (reversible)

BMI = body mass index; Pedigree = diabetes pedigree function; Ca = number of major vessels coloured by fluoroscopy; Thal = thallium stress test result; Oldpeak = exercise-induced ST depression.

4.5 Case Study: Redundant Features in WDBC

Figure 4 shows the fold-to-fold distribution of importance for the features that appear in the top three of any of the four equally accurate models on WDBC. The linear and kernel models spread their importance thinly across many correlated size, shape and texture descriptors, each with small importance and overlapping distributions, so that their ordering changes from fold to fold. RF and GB concentrated importance on the size-related features worst area, worst perimeter and worst concave points, but because these features are nearly collinear (Spearman $|\rho| > 0.9$), which of them receives the credit varies between fits. Both patterns are consequences of redundancy: when several features carry the same information, permutation importance is either diluted across them or assigned somewhat arbitrarily among them [14], [15].

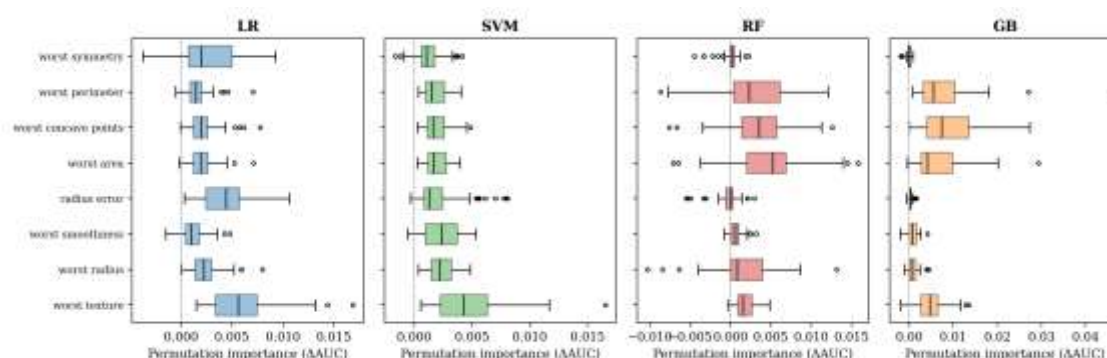


Figure 4. Distribution of Permutation Importance (Decrease in Test AUC) Across 50 Fits on WDBC for the Features Ranked in the Top Three by Any of the Four Equally Accurate Models. the Dashed Line Marks Zero Importance

4.6 Effect of Training-Set Size

Table 8 and Figure 5 report AUC and stability when the training folds were subsampled to 25%, 50% and 100% of their size. For LR, SVM, RF and GB, reducing the training data to one quarter lowered mean AUC by at most 0.030 (GB on Pima) and typically by less than 0.02. Explanation stability, by contrast, fell sharply: for LR it dropped from 0.90 to 0.68 on Pima and from 0.55 to 0.26 on Cleveland, and for SVM from 0.85 to 0.65 on Pima. In other words, discrimination saturated at much smaller sample sizes than

explanation stability. RF on WDBC was the exception, with low stability at all sizes (0.25–0.36), consistent with the redundancy effect described in Section 4.5.

Table 8. AUC and Explanation Stability ($\bar{\rho}$) When Each Training Fold is Subsampled to 25%, 50% or 100% of Its Size (5 × 5-Fold Cross-Validation, 25 Fits per Cell)

Dataset	Model	AUC 25%	AUC 50%	AUC 100%	$\bar{\rho}$ 25%	$\bar{\rho}$ 50%	$\bar{\rho}$ 100%
WDBC	LR	0.993	0.995	0.995	0.35	0.41	0.56
	SVM	0.990	0.994	0.995	0.48	0.47	0.59
	RF	0.988	0.989	0.990	0.36	0.25	0.30
	GB	0.987	0.991	0.992	0.31	0.32	0.46
	MLP†	0.316	0.967	0.982	0.81	0.57	0.41
Pima	LR	0.822	0.832	0.836	0.68	0.77	0.90
	SVM	0.806	0.819	0.830	0.65	0.78	0.85
	RF	0.811	0.822	0.831	0.70	0.77	0.83
	GB	0.784	0.801	0.814	0.62	0.68	0.78
	MLP†	0.480	0.494	0.692	0.88	0.89	0.69
Cleveland	LR	0.874	0.893	0.899	0.26	0.36	0.55
	SVM	0.878	0.886	0.895	0.33	0.40	0.47
	RF	0.881	0.886	0.897	0.39	0.49	0.54
	GB	0.871	0.862	0.882	0.33	0.38	0.49
	MLP†	0.668	0.790	0.859	0.63	0.55	0.52

†Degenerate MLP fits: at reduced sample sizes the MLP frequently produced near-constant predictions (AUC ≤ 0.5 on WDBC at 25% and on Pima at 25–50%), for which all importances are close to zero and the resulting rankings are spuriously consistent.

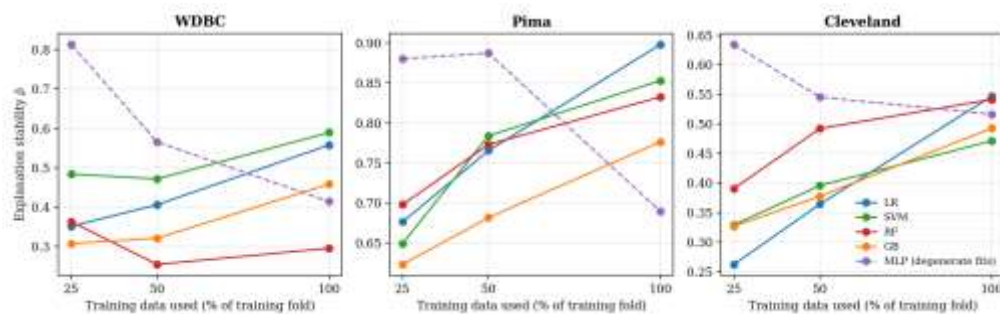


Figure 5. Explanation Stability ($\bar{\rho}$) as a Function of the Fraction of Each Training Fold Used for Fitting. the Dashed MLP Curve Reflects Degenerate Fits at Small Sample Sizes and Should Not be Read as Genuine Stability

The MLP curves illustrate a failure mode of stability metrics. At 25% of the training data, the MLP reached an AUC of only 0.316 on WDBC and 0.480 on Pima, yet produced the highest stability values in the experiment (0.81 and 0.88). When a model makes nearly constant predictions, permuting any feature changes nothing, all importances are close to zero, and many fits share the same trivial ranking. High stability is therefore a necessary but not sufficient condition for a trustworthy explanation and must always be interpreted together with the performance of the model being explained.

4.7 Discussion

4.7.1 Principal Findings

This study shows that equal accuracy does not imply equally reliable explanations. On WDBC, four classifiers that were statistically indistinguishable in AUC produced permutation-importance rankings whose stability differed almost two-fold and whose leading features did not coincide. Across datasets, stability was governed more by the structure of the data, in particular the number of redundant features, and by the amount of training data than by the choice of model. The positive association between accuracy and stability across all five models ($\rho = 0.60\text{--}0.90$) is driven mainly by the poorly performing MLP; among competitive models, accuracy was a weak guide to explanation reliability. These findings provide empirical, clinically relevant evidence for the concerns raised in the Rashomon and underspecification literature [4], [6], [7], [9].

4.7.2 Why Explanations Become Unstable

Three mechanisms account for the observed instability. First, feature redundancy: when predictors are strongly correlated, as for the radius, perimeter and area descriptors of WDBC, a model can obtain the same information from any member of a correlated group, so credit is diluted or assigned arbitrarily, and permutation additionally evaluates the model on unrealistic feature combinations [14], [15]. Second, model structure: the regularised linear and kernel models distribute weight smoothly across correlated inputs, whereas tree ensembles select among them through discrete splits, so small changes in the training sample change which feature is split upon. This explains why RF combined low whole-ranking stability with moderate top-5 stability. Third, estimation noise: importance is computed on test folds of 60–154 observations, and for weak features its sampling error is comparable to its magnitude, as indicated by negative-importance shares of 15–33%. The training-size experiment shows that this third component shrinks slowly: explanations continued to stabilise long after AUC had saturated.

4.7.3 Implications for Practice

The results suggest four practical recommendations for studies that report feature importance, complementary to existing reporting guidance such as TRIPOD+AI [19]. (i) Importance should be computed on held-out data within a resampling scheme and reported with a stability measure, such as $\bar{\rho}$ and top-k Jaccard, not from a single fit. (ii) When several models are equally accurate, explanation stability, together with calibration [20], [25], is a legitimate criterion for choosing among them; on WDBC, for instance, LR and SVM offer the same discrimination as the tree ensembles with better calibration and more stable explanations, in line with the argument for simpler models in high-stakes settings [7], [12]. (iii) Correlated features should be grouped or handled with conditional or grouped importance before their individual ranks are interpreted [14], [16]. (iv) Stability must be read alongside performance, because degenerate models can appear highly stable. Finally, statements that a model “identified” a clinical risk factor should be reserved for features that are consistently important across refits and across model families, such as glucose in Pima and Ca in Cleveland.

4.7.4 Limitations

Several limitations should be noted. The study used three public datasets of modest size; larger, higher-dimensional clinical cohorts may behave differently, although the redundancy mechanism is expected to become stronger rather than weaker. Hyperparameters were fixed rather than tuned to isolate sampling variability, which disadvantaged the MLP in particular; tuning within each fold would probably improve its performance but would also add a further source of explanation variability. Only permutation importance was studied; SHAP [23] and other attribution methods may show different stability profiles and should be examined with the same protocol. The statistical tests on stability rely on leave-one-out values that are not strictly independent, and the Friedman tests on resampled folds are anti-conservative; the conclusions therefore rest mainly on effect sizes and on the corrected t-tests for AUC. Finally, the study measures the reproducibility of explanations, not their causal or clinical validity, which requires domain expertise and external evidence.

5. CONCLUSION

Permutation feature importance is a convenient and widely used way to explain clinical ML models, but its reproducibility cannot be taken for granted. In a controlled benchmark of five classifier families on three clinical datasets with 50 refits per configuration, models with statistically indistinguishable discrimination produced explanations whose stability differed almost two-fold and whose most important features disagreed. Stability decreased with feature redundancy and with smaller training sets and, unlike AUC, continued to improve well beyond the sample size at which discrimination had saturated. Degenerate models, conversely, could appear spuriously stable. Accuracy is therefore not a certificate of explanation reliability. Studies that report feature importance should quantify its stability across refits, handle correlated features explicitly, and interpret explanations only in conjunction with model performance and calibration. Future work will extend the protocol to SHAP and other attribution methods, to tuned and larger models, and to external validation cohorts.

Author Contribution Statement

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Surjeet Dalal	✓	✓	✓		✓				✓		✓			
Mohd Akbar		✓		✓		✓		✓		✓		✓		

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

Acknowledgments

The authors thank the creators and maintainers of the Wisconsin diagnostic breast cancer, Pima Indians diabetes and Cleveland heart disease datasets, and the developers of the open-source scientific Python ecosystem, whose work made this study possible.

Funding Information

This research has received no external funding.

Conflict of Interest Statement

No conflict of interest.

Informed Consent

Not Applicable.

Ethical Approval

Not Applicable. The study used only publicly available, de-identified datasets.

Data Availability

This study draws on publicly available data: the Wisconsin diagnostic breast cancer dataset (distributed with scikit-learn), the Pima Indians diabetes dataset and the Cleveland heart disease dataset. No new datasets were created. The analysis code is available from the corresponding author on reasonable request.

REFERENCES

- [1] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018. doi.org/10.1109/ACCESS.2018.2870052
- [2] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, 2021. doi.org/10.1109/TNNLS.2020.3027314
- [3] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. doi.org/10.1023/A:1010933404324
- [4] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," *Journal of Machine Learning Research*, vol. 20, no. 177, pp. 1–81, 2019. jmlr.org/papers/v20/18-760.html
- [5] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. jmlr.org/papers/v12/pedregosa11a.html
- [6] L. Breiman, "Statistical modeling: The two cultures," *Statistical Science*, vol. 16, no. 3, pp. 199–231, 2001. doi.org/10.1214/ss/1009213726
- [7] L. Semenova, C. Rudin, and R. Parr, "On the existence of simpler machine learning models," in *Proc. 2022 ACM Conf. Fairness, Accountability, and Transparency*, 2022, pp. 1827–1858. doi.org/10.1145/3531146.3533232
- [8] C. Marx, F. Calmon, and B. Ustun, "Predictive multiplicity in classification," in *Proc. 37th Int. Conf. Machine Learning*, PMLR, vol. 119, 2020, pp. 6765–6774. proceedings.mlr.press/v119/marx20a.html
- [9] A. D'Amour et al., "Underspecification presents challenges for credibility in modern machine learning," *Journal of Machine Learning Research*, vol. 23, no. 226, pp. 1–61, 2022. jmlr.org/papers/v23/20-1335.html
- [10] S. Krishna, T. Han, A. Gu, J. Pombra, S. Jabbari, S. Wu, and H. Lakkaraju, "The disagreement problem in explainable machine learning: A practitioner's perspective," *arXiv preprint arXiv:2202.01602*, 2022. doi.org/10.48550/arXiv.2202.01602
- [11] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *The Lancet Digital Health*, vol. 3, no. 11, pp. e745–e750, 2021. [doi.org/10.1016/S2589-7500\(21\)00208-9](https://doi.org/10.1016/S2589-7500(21)00208-9)
- [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. doi.org/10.1038/s42256-019-0048-x
- [13] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinformatics*, vol. 8, p. 25, 2007. doi.org/10.1186/1471-2105-8-25
- [14] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinformatics*, vol. 9, p. 307, 2008. doi.org/10.1186/1471-2105-9-307
- [15] G. Hooker, L. Mentch, and S. Zhou, "Unrestricted permutation forces extrapolation: Variable importance requires at least one more model, or there is no free variable importance," *Statistics and Computing*, vol. 31, no. 6, p. 82, 2021. doi.org/10.1007/s11222-021-10057-z
- [16] C. Molnar et al., "General pitfalls of model-agnostic interpretation methods for machine learning models," in *xxAI – Beyond Explainable AI*, *Lecture Notes in Computer Science*, vol. 13200. Cham: Springer, 2022, pp. 39–68. doi.org/10.1007/978-3-031-04083-2_4
- [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. doi.org/10.1007/s10115-006-0040-8
- [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. jmlr.org/papers/v18/17-514.html

- [19] G. S. Collins et al., "TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods," *BMJ*, vol. 385, p. e078378, 2024. doi.org/10.1136/bmj-2023-078378
- [20] 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, p. 230, 2019. doi.org/10.1186/s12916-019-1466-7
- [21] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," *arXiv preprint arXiv:1806.08049*, 2018. doi.org/10.48550/arXiv.1806.08049
- [22] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144. doi.org/10.1145/2939672.2939778
- [23] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems* 30, 2017, pp. 4765–4774. doi.org/10.48550/arXiv.1705.07874
- [24] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017. doi.org/10.48550/arXiv.1702.08608
- [25] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proc. 22nd Int. Conf. Machine Learning*, 2005, pp. 625–632. doi.org/10.1145/1102351.1102430
- [26] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Machine Learning*, PMLR, vol. 70, 2017, pp. 1321–1330. proceedings.mlr.press/v70/guo17a.html
- [27] C. Nadeau and Y. Bengio, "Inference for the generalization error," *Machine Learning*, vol. 52, no. 3, pp. 239–281, 2003. doi.org/10.1023/A:1024068626366
- [28] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006. jmlr.org/papers/v7/demsar06a.html
- [29] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," in *Proc. SPIE 1905, Biomedical Image Processing and Biomedical Visualization*, 1993, pp. 861–870. doi.org/10.1117/12.148698
- [30] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in *Proc. Annu. Symp. Computer Applications in Medical Care*, 1988, pp. 261–265. www.ncbi.nlm.nih.gov/pmc/articles/PMC2245318/
- [31] R. Detrano et al., "International application of a new probability algorithm for the diagnosis of coronary artery disease," *The American Journal of Cardiology*, vol. 64, no. 5, pp. 304–310, 1989. [doi.org/10.1016/0002-9149\(89\)90524-9](https://doi.org/10.1016/0002-9149(89)90524-9)
- [32] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995. doi.org/10.1007/BF00994018
- [33] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001. doi.org/10.1214/aos/1013203451
- [34] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems* 30, 2017, pp. 3146–3154. papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree
- [35] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986. doi.org/10.1038/323533a0
- [36] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982. doi.org/10.1148/radiology.143.1.7063747
- [37] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, 1950. [doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2)
- [38] M. P. Naeini, G. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using Bayesian binning," in *Proc. 29th AAAI Conf. Artificial Intelligence*, 2015, pp. 2901–2907. doi.org/10.1609/aaai.v29i1.9602

- [39] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the American Statistical Association*, vol. 32, no. 200, pp. 675–701, 1937. doi.org/10.1080/01621459.1937.10503522
- [40] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945. doi.org/10.2307/3001968
- [41] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979. www.jstor.org/stable/4615733

How to Cite: Surjeet Dalal, Mohd Akbar. (2026). Equally accurate, differently explained: Stability of permutation feature importance across machine learning classifiers on clinical tabular data. *Journal of Artificial Intelligence, Machine Learning and Neural Network (JAIMLNN)*, 6(2), 15-31. <https://doi.org/10.55529/jaimlnn.62.15.31>

BIOGRAPHIES OF AUTHORS

	Prof. (Dr.) Surjeet Dalal, is a Professor and Head of the Department of Computer Science and Engineering at Amity School of Engineering and Technology, Amity University Gurugram, Haryana, India, and a Visiting Professor at the City Institute of Higher Learning, Canada. His research concerns artificial intelligence and machine learning for clinical and health-care applications, including predictive modelling, decision support and the interpretability of models used in medical settings.
	Dr. Mohd Akbar is Senior Teaching Staff in the Department of Computing and Information Sciences, School of Computing and Information Sciences, University of Technology and Applied Sciences, Muscat, Sultanate of Oman. His research concerns machine learning on tabular and structured data, including model interpretability and the behaviour of feature importance measures. He is interested in how explanations of model behaviour are produced, compared and reported in applied work.