MACHINE LEARNING

Explainable Machine Learning for Breast Tumour Classification: A Reproducible Wisconsin Diagnostic Dataset Study

Priyanshu Raghav, Ritesh Kumar

Department of Artificial Intelligence & Data Science, Dr. Akhilesh Das Gupta Institute of Professional Studies, Delhi, India

Published November 2026

Abstract

Background: Medical-AI benchmarks require transparent model selection, genuine evaluation outputs and qualified explanations. Objective: To compare three machine-learning procedures for benign/malignant breast tumour classification and explain the development-selected family. Methods: The public WDBC dataset supplied 569 records and 30 FNA-derived numerical predictors. A stratified split reserved 455 development and 114 internal holdout records. Five-outer/three-inner nested validation selected the family by ROC-AUC; fold-local scaling and frozen final settings prevented test-driven tuning. Permutation reliance was evaluated on development validation folds only. Results: RBF SVM led on mean nested AUC (0.9961 ± 0.0047). On the holdout it achieved 113/114 correct predictions, 97.62% malignant sensitivity, 100.00% observed specificity and AUC 0.9970 (stratified-bootstrap 95% interval 0.9891–1.0000). It missed one malignant record. Logistic Regression and Random Forest achieved accuracies of 96.49% and 97.37%. Paired AUC-difference intervals included zero. Worst texture and measurement-error summaries led SVM permutation reliance, with strong predictor correlations limiting interpretation. Conclusion: The findings demonstrate reproducible internal benchmark discrimination, not established clinical superiority or patient-facing readiness. Source data, code, predictions and exact fixed-protocol rerun checks accompany the study.

Keywords: Breast tumour classification; Explainable AI; Nested cross-validation; Support vector machine; WDBC

1 Introduction

Explainable medical artificial intelligence (AI) requires inspectable evaluation as well as readable model outputs. The Wisconsin Diagnostic Breast Cancer (WDBC) dataset provides a tractable setting: 30 measurements derived from fine-needle aspirate (FNA) images describe cell nuclei in 569 breast-mass records labelled benign or malignant [1,2]. A transparent analysis can compare model families and explain feature reliance while making the boundary between benchmark performance and clinical evidence explicit.

This study conducts a new computational comparison of Logistic Regression, radial-basis-function support vector machine (RBF SVM) and Random Forest. The objective is not to introduce a new algorithm or claim a clinical screening service. It is to determine which specified procedure leads on development-set ROC-AUC, quantify the frozen models' internal holdout performance and examine the selected family's reliance on correlated features. All numerical performance findings below come from executed code rather than being borrowed from published experiments.

2 Background and related methods

The original WDBC measurements represent ten families of nuclear characteristics, each summarised by its mean, standard error and worst values. The source defines worst as the mean of the three largest measurements. These are numeric predictors inherited from an existing measurement pipeline; this study does not train an image extractor or evaluate specimen acquisition [1,2]. Logistic Regression, SVM and Random Forest supply complementary linear, nonlinear kernel and tree-ensemble modelling strategies [3–5].

Two methodological issues motivate the design. First, estimating preprocessing or selecting settings from evaluation observations can create leakage. Second, the best tuning score can be optimistic because it was used to choose settings. Nested cross-validation places parameter search inside an outer evaluation loop, while a separate holdout limits reliance on development comparisons [6,7]. Choosing a family from outer means still introduces development-stage selection, so those means are not treated as definitive clinical evidence about the winner.

Permutation importance measures model reliance through a change in predictive score after shuffling an input. It is model- and dataset-dependent, and correlated predictors can substitute for each other [8]. TRIPOD+AI supports transparent reporting of data, methods and limitations, but reporting guidance does not itself establish a model's safety or applicability to a new clinical population [9]. The present work uses these principles without claiming formal checklist compliance.

3 Data and experimental protocol

3.1 Source, audit and target

The official UCI WDBC archive and data dictionary were retained. The audit confirmed 569 records, 357 benign labels, 212 malignant labels, 30 finite numeric predictors and no missing predictor cells, repeated identifiers or exactly duplicated predictor rows. The raw measurements also matched scikit-learn's bundled dataset numerically. Identifiers and diagnosis were excluded from the predictor matrix. Malignant was explicitly recoded to 1 and benign to 0, reversing the bundled dataset's numeric target convention [1,3].

A stratified 80/20 split with seed 42 produced 455 development records, comprising 170 malignant and 285 benign observations, and 114 holdout records, comprising 42 malignant and 72 benign observations. The split manifest preserves membership and public source identifiers. Record identifiers do not overlap across partitions. This is internal random-split evaluation within one historical dataset, not an external-hospital assessment.

Table 1. Dataset and evaluation design.

Item	Executed specification
Data	569 WDBC records; 30 FNA-derived numeric predictors
Target	Malignant = 1; benign = 0
Development / holdout	455 (170 M, 285 B) / 114 (42 M, 72 B)
Selection	5 outer / 3 inner folds; ROC-AUC; development only
Final fitting	5-fold retuning on development; parameters frozen
Explanation	30 shuffles per feature in each development outer validation fold

Official data; all fitting and evaluation details are recorded in protocol.json. The holdout is internal to WDBC, not external clinical validation.

3.2 Model selection and frozen evaluation

Five stratified development outer folds used shuffle seed 142. Each held out 91 records while an inner three-fold ROC-AUC grid search used the remaining 364 records. Inner seeds were 242 plus outer fold number. The prespecified grids contained 10 Logistic Regression, 18 SVM and eight Random Forest configurations. The family with the largest mean outer ROC-AUC was selected; exact ties would follow the fixed family order. A local protocol file recorded this plan before fitting, but the study was not externally preregistered.

Logistic Regression and SVM used StandardScaler inside their pipelines and every cross-validation fit. Random Forest used unscaled features and 300 trees. Every family retained all 30 predictors. No imputation, resampling or feature selection was performed. Each family was subsequently retuned using five stratified folds on all development records with seed 342; final settings and the family choice were frozen before holdout scoring. Optimisation scores from this final tuning stage are not reported as unbiased evaluation estimates [6,7].

The default predict rules were retained. Logistic Regression and Random Forest generally use a malignant probability boundary of 0.5; an exact forest tie follows class order. SVM uses a zero signed-margin boundary. Its scores were not converted to probabilities. A most-frequent-class baseline was also fitted on development records. No threshold or hyperparameter was changed in response to holdout mistakes.

3.3 Metrics and uncertainty

The primary metric was ROC-AUC. Secondary measures were accuracy, malignant precision, sensitivity, specificity, F1 and balanced accuracy. True positive and false negative always refer to malignant records. Reported outer-fold standard deviations are descriptive variation across five overlapping training procedures, not confidence intervals or five independent patient cohorts.

Holdout AUC intervals used 2,000 class-stratified bootstrap resamples with seed 2026 and frozen scores. Each resample contained 42 malignant and 72 benign observations. The percentile interval used the 2.5th and 97.5th percentiles. Sensitivity, specificity and accuracy intervals used Wilson's method [10]. Exploratory paired AUC differences used the same bootstrap indices. All intervals condition on the fitted models and observed source and omit retraining and external-population uncertainty.

4 Results

4.1 Development selection and final settings

Mean nested outer ROC-AUC was 0.9956 with standard deviation 0.0069 for Logistic Regression, 0.9961 with standard deviation 0.0047 for RBF SVM, and 0.9900 with standard deviation 0.0104 for Random Forest. The SVM was therefore selected by the fixed criterion. Its approximately 0.00052 AUC-unit mean advantage over Logistic Regression was small. Logistic Regression instead had the highest mean outer accuracy, 97.80%, versus 96.92% for SVM and 95.82% for Random Forest. The study does not retrospectively redefine the selection objective.

The final development-only settings were C = 1 and no class weighting for Logistic Regression; C = 10, gamma = 0.01 and balanced weights for RBF SVM; and unrestricted depth, square-root predictor sampling and minimum leaf size one for the 300-tree forest. Individual outer-fold choices differed because their training sets were smaller. Complete choices and grid scores are retained in the reproducibility files.

4.2 Holdout metrics and errors

The models correctly classified 110, 113 and 111 of 114 holdout records, respectively. Logistic Regression produced 39 true positives, three false negatives, 71 true negatives and one false positive. SVM produced 41 true positives, one false negative, 72 true negatives and no false positives. Random Forest produced 39 true positives, three false negatives, 72 true negatives and no false positives. The majority baseline attained 63.16% accuracy but zero malignant sensitivity and ROC-AUC 0.5000.

Table 2. Development ROC-AUC and measured holdout performance.

Metric	Logistic Regression	RBF SVM	Random Forest
CV AUC mean ± SD	0.9956 ± 0.0069	0.9961 ± 0.0047	0.9900 ± 0.0104
Test accuracy	96.49%	99.12%	97.37%
Test sensitivity	92.86%	97.62%	92.86%
Test specificity	98.61%	100.00%	100.00%
Test F1	0.9512	0.9880	0.9630
Test AUC	0.9960	0.9970	0.9944

Executed results; malignant is positive. Development values are nested outer-fold means and sample SDs, not confidence intervals. All test metrics use the same 114 records and default predict rules.

Holdout AUCs were 0.9960, 0.9970 and 0.9944 for Logistic Regression, SVM and Random Forest. Stratified-bootstrap 95% intervals were 0.9861–1.0000, 0.9891–1.0000 and 0.9841–1.0000. Figure 1 uses saved probabilities for Logistic Regression and Random Forest and signed margins for SVM; no calibration comparison is implied. Strong discrimination within this benchmark does not specify a clinically suitable action threshold.

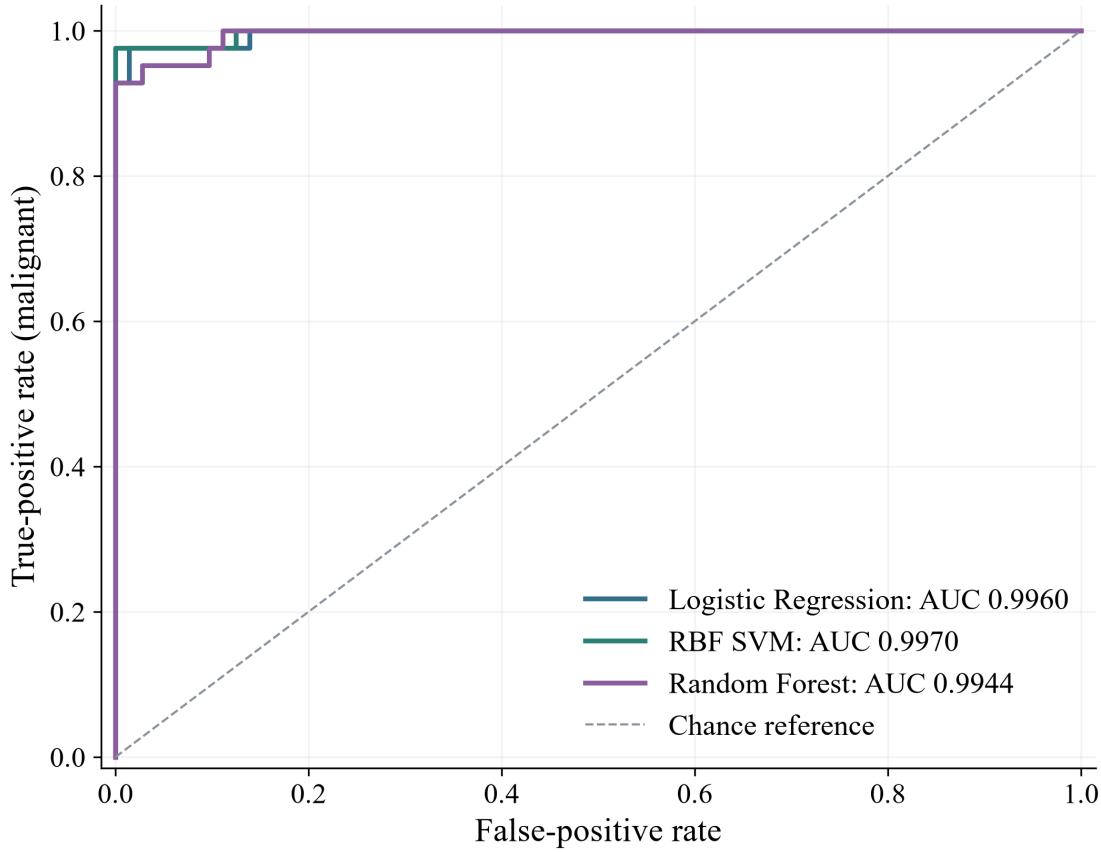

Figure 1. Internal holdout ROC curves from frozen model scores. SVM uses signed margins, not probabilities; the curves share 114 records.

For the selected SVM, sensitivity was 41/42 = 97.62% with Wilson 95% interval 87.68–99.58%, and specificity was 72/72 = 100.00% with interval 94.93–100.00%. Accuracy was 113/114 = 99.12%, with interval 95.20–99.84%. Zero observed false positives therefore does not establish universally perfect specificity. The single malignant false negative remains visible in Figure 2.

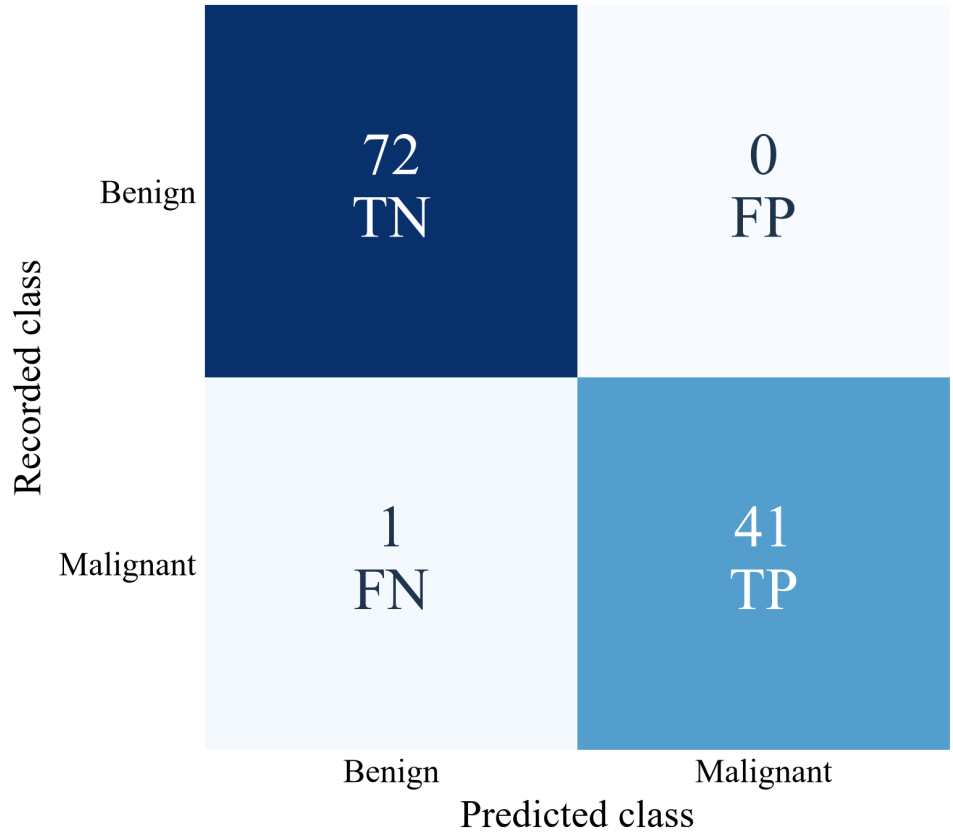

Figure 2. Selected SVM confusion matrix: TN 72, FP 0, FN 1, TP 41. One of 42 malignant holdout records was missed.

4.3 Paired differences and error review

The SVM-minus-Logistic-Regression AUC difference was 0.00099, with paired percentile interval −0.00066 to 0.00397. The SVM-minus-Random-Forest difference was 0.00265, with interval −0.00331 to 0.01058. Both include zero. They do not prove equivalence, but they do not support a clear AUC-superiority claim on this holdout. No multiplicity-adjusted hypothesis test or clinical non-inferiority analysis was performed.

All three models missed the same malignant record, T017 in the retained test order. Its SVM margin was −1.10665, illustrating that the common error was not simply an infinitesimal boundary crossing. Logistic Regression's only false positive, T005, had an uncalibrated malignant probability estimate of 0.50916, while missed malignant record T113 had estimate 0.49161. These examples are descriptive post-evaluation error review, not evidence for a threshold adjustment. No clinical re-adjudication of labels occurred.

5 Feature-based explanation

The selected SVM family was explained using its five development outer-fold models and their corresponding validation records. Each of 30 features was shuffled 30 times per fold, giving 4,500 score-decrease measurements. The mean of five fold-level average decreases defines the reported importance, and their sample standard deviation summarises fold variability. The holdout was not used for explanation-driven feature selection or model revision [8].

Worst texture had the largest mean decrease in AUC, 0.00525, followed by radius error at 0.00283 and perimeter error at 0.00181. The first two had positive fold-level means in all five folds; perimeter error was positive in four. Table 3 and Figure 3 describe reliance of the fitted family under this score, not causal diagnostic relevance or individual-patient attribution.

Table 3. Leading development-only SVM permutation importances.

Feature	Mean AUC decrease	SD across fold means	Positive folds
worst texture	0.00525	0.00252	5/5
radius error	0.00283	0.00241	5/5
perimeter error	0.00181	0.00174	4/5
compactness error	0.00172	0.00252	3/5
worst radius	0.00168	0.00054	5/5

Thirty shuffles per feature per outer validation fold. Values describe model reliance, not causal effects. SD is not a confidence interval; test data did not select these features.

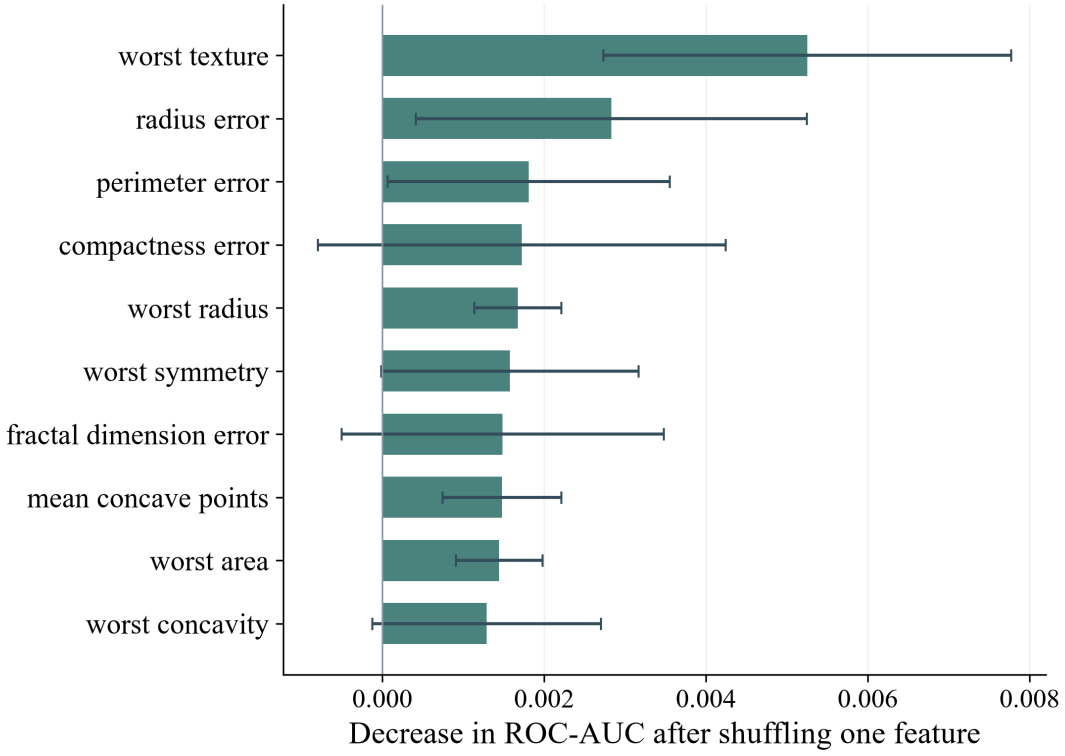

Figure 3. Development-validation SVM permutation importance. Bars are mean AUC decreases; error ranges are SD across five fold means, not confidence intervals.

Development Spearman correlation was 0.9996 between mean radius and mean area, 0.9976 between mean radius and mean perimeter, and 0.9987 between worst radius and worst area. Radius error and perimeter error correlated at 0.9572. Such redundancy lets the model retain similar information when a single measurement is shuffled. Low individual permutation importance consequently cannot be interpreted as intrinsic irrelevance. Error bars are standard deviations across fold means, not 95% intervals; rankings do not indicate feature-effect direction [8].

6 Discussion and limitations

The experiment demonstrates that high internal discrimination can coexist with a meaningful malignant error, wide sensitivity uncertainty and uncertain differences between model families. It also shows why a development selection metric should be retained: the SVM led on mean AUC, whereas Logistic Regression led on mean development accuracy. An explanation plot adds information about fitted model behaviour but cannot remove these evaluation limits.

The data are a small historical benchmark with already-extracted features. The study does not evaluate raw FNA images, acquisition failures, missingness, temporal drift or multicentre transportability. The available fields do not support a demographic fairness audit. A random holdout from one source is not external validation, and no claim about a particular hospital or Indian clinical population is justified. Modest unequal-sized grids compare specified procedures, not each family's maximum attainable performance [1,7,9].

Probability calibration, prospective decision utility, patient outcomes, cost effectiveness, local explanation validity and causal analysis were not evaluated. Conditional bootstrap intervals omit development-sample uncertainty. Family selection from development outer means retains some selection optimism; their standard deviations are not inferential intervals. The present result is a reproducible computational benchmark, not evidence that the supplied classifier should guide patient care.

7 Conclusion

A leakage-controlled WDBC comparison selected RBF SVM using development nested ROC-AUC and obtained 113 correct predictions among 114 internal holdout records, with one malignant false negative and AUC 0.9970. Small paired AUC differences from Logistic Regression and Random Forest remained uncertain. Development-only permutation analysis identified worst texture and measurement-error summaries as prominent sources of fitted SVM reliance, with strong feature correlations limiting individual rankings.

The contribution is an auditable experiment linking official data, fixed partitions, executed models, saved predictions and qualified explanations. It is not a new algorithm or a clinical validation. Independent evaluation with compatible measurements, training-only calibration and clinically governed decision rules would be required before considering patient-facing research.

8 Data, code and ethics statement

The public UCI WDBC dataset is distributed under CC BY 4.0 [1]. Source data, attribution, executable Python, protocol, split manifest and measured result artifacts accompany the report. A second unchanged execution reproduced the selection, metrics, predictions and importance summaries exactly; it is not a new independent sample. No patients were recruited, no new clinical data were collected and no institutional ethics approval, clinical deployment or publication status is claimed.

References

[1] UCI Machine Learning Repository. (n.d.). Breast Cancer Wisconsin (Diagnostic) [Dataset and data dictionary]. https://doi.org/10.24432/C5DW2B

[2] Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993). Nuclear feature extraction for breast tumor diagnosis. *Proceedings of SPIE*, 1905, 861–870. https://doi.org/10.1117/12.148698

[3] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

[4] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297. https://doi.org/10.1007/BF00994018

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

[6] Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7, 91. https://doi.org/10.1186/1471-2105-7-91

[7] Scikit-learn developers. (2025). Cross-validation: evaluating estimator performance. *Scikit-learn* 1.6.1 documentation. https://scikit-learn.org/1.6/modules/cross_validation.html

[8] Scikit-learn developers. (2025). Permutation feature importance. *Scikit-learn* 1.6.1 documentation. https://scikit-learn.org/1.6/modules/permutation_importance.html

[9] Collins, G. S., Moons, K. G. M., Dhiman, P., et al. (2024). TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385, e078378. https://doi.org/10.1136/bmj-2023-078378

[10] Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158), 209–212. https://doi.org/10.2307/2276774