

Classifying Breast Tumours from Cell-Nucleus Measurements with Machine Learning: A Leakage-Free Comparison of Feature Sets for AI-Assisted Decision Support

KARTIK PHOPASE

`kartik.phopse_24uds@sanjivani.edu.in`

sanjivani university <https://orcid.org/0009-0002-4168-179X>

Research Article

Keywords: Machine Learning, Breast Cancer Diagnosis, Medical Decision Support, Feature Subsets, Permutation Importance, Model Evaluation

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11491091/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Classifying Breast Tumours from Cell-Nucleus Measurements with Machine Learning: A Leakage-Free Comparison of Feature Sets for AI-Assisted Decision Support

Kartik Phopase

ORCID:  0009-0002-4168-179X

Department of Artificial Intelligence and Data Science

Sanjivani University, Maharashtra, India

Preprint, October 2026

ABSTRACT

Abstract—Computer-aided diagnosis tools need classifiers that separate malignant from benign breast masses reliably, and that remain trustworthy when only part of the available measurements can be obtained. This paper studies that problem as a binary classification task on the public Wisconsin Diagnostic Breast Cancer dataset of 569 fine-needle-aspirate samples, where the positive class is malignant (37.3%). Four machine learning models (logistic regression, RBF support vector machine, random forest, gradient boosting) and a majority-class baseline are compared with repeated stratified 5×10 -fold cross-validation on a 455-sample training split and a single evaluation on a 114-sample held-out test split. Two scenarios are examined: (A) a compact scenario using only the 10 “mean” nucleus measurements, and (B) a full scenario using all 30 measurements (mean, standard error, worst value). All models reach a cross-validated ROC-AUC of 0.98–0.99 in both scenarios. In scenario A the selected gradient boosting model achieves a held-out accuracy of 0.947 and AUC of 0.979 (95% CI 0.947–0.998); in scenario B the selected SVM achieves accuracy 0.965 and AUC of 0.995 (95% CI 0.982–1.000). The two scenarios differ by less than the uncertainty of the evaluation, so the added features give at most a small, statistically unconfirmed gain. Permutation importance concentrates on mean area and mean concave points in scenario A and is diffuse in scenario B because many features are strongly correlated. The results are a methodological benchmark, not evidence of clinical readiness.

Index Terms—Machine Learning, Breast Cancer Diagnosis, Medical Decision Support, Feature Subsets, Permutation Importance, Model Evaluation

I. INTRODUCTION

Machine learning is increasingly used in medical decision support, including cytology and radiology workflows where a model estimates the probability that a lesion is malignant. For fine-needle aspiration (FNA) of breast masses, a well-known line of work extracts numerical descriptors of cell nuclei from digitised images and trains classifiers on them [1, 2, 3]. These descriptors are cheap to compute, but a deployed system may not always have all of them, or may prefer a simpler model that relies on fewer inputs.

This paper therefore asks a focused empirical question: how much does classification performance change when a model uses only the 10 mean nucleus measurements compared with the full set of 30 measurements, and do conclusions depend on the choice among standard models? The main contributions are:

- A reproducible comparison of four machine learning models and a baseline on a public diagnostic dataset.
- Two feature scenarios—compact (mean values) and full (mean, standard error, worst)—evaluated under one leakage-free protocol.
- An honest account of uncertainty: bootstrap intervals on a held-out set, the size of fold-to-fold variation, and a discussion of why permutation importance is hard to interpret when features are strongly correlated.

II. PROBLEM STATEMENT

A. Unreported Evaluation Uncertainty

Studies on this dataset frequently report accuracies above 95% for many different models, often from a single run or from cross-validation used both to tune and to report results [7]. When every model scores near the ceiling, rankings between

models are easily over-interpreted.

B. Need for Guidance on Feature Availability

Decision-support systems may operate with a reduced set of measurements. It is unclear how much of the achievable performance is retained with the compact feature set, and how much of any remaining gap is larger than the noise in the evaluation.

III. RESEARCH OBJECTIVES

1. To develop and compare machine learning models that classify a breast mass as malignant or benign.
2. To evaluate prediction quality in a compact scenario (10 mean features) and a full scenario (30 features), to quantify the uncertainty of the comparison, and to identify the most influential attributes.

IV. LITERATURE REVIEW

A. Computer-Aided Diagnosis from FNA Images

Wolberg and Mangasarian applied a multisurface pattern-separation method to breast cytology data for diagnosis [2]. Street et al. described the extraction of nuclear features from digitised FNA images, which became the basis of the dataset used here [1]. Mangasarian et al. later addressed both diagnosis and prognosis of breast cancer using linear programming [3]. The dataset is distributed through the UCI Machine Learning Repository [4].

B. Machine Learning in Cancer Research

Kourou et al. reviewed machine learning applications in cancer prognosis and prediction and noted the importance of validation and of the quality of the underlying data [5]. Rudin argued that, for high-stakes decisions, interpretable models should be preferred over post-hoc explanations of black-box

TABLE 1. INPUT VARIABLES BY STATISTIC

Group	Examples	Count
Mean (scenarios A, B)	mean radius, mean area	10
Standard error (B only)	radius error, area error	10
Worst value (B only)	worst radius, worst texture	10

models [6].

C. Evaluation Methodology

Cawley and Talbot showed that using the same data for model selection and for performance estimation produces optimistic results [7]. Hooker et al. discussed the limitations of permutation-based importance when features are dependent [8]. The calibration of predicted probabilities can be summarised by the Brier score [9].

V. RESEARCH GAP

Prior work reports high accuracy on this dataset, but often without a held-out evaluation separate from model selection, and without confidence intervals. This study addresses the narrower gap of comparing a compact and a full feature set under a single, leakage-free evaluation protocol with uncertainty estimates, and of stating clearly what the data can and cannot support.

VI. RESEARCH METHODOLOGY

Fig. 1 summarises the pipeline.

A. Dataset

The data are the Wisconsin Diagnostic Breast Cancer (WDBC) dataset [4, 1]: 569 FNA samples, of which 357 are benign and 212 malignant, with no missing values. For each sample, ten real-valued nucleus measurements (radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension) are summarised by three statistics—the mean, the standard error and the worst (largest) value—giving 30 numeric features (Table 1). The data were loaded through scikit-learn.

B. Target and Scenarios

The target is malignant = 1 and benign = 0; 37.3% of samples are malignant, so the majority-class baseline predicts “benign” and has accuracy 0.626 on the test set. Malignant is treated as the positive class, so precision, recall and F1 describe malignancy detection. Scenario A (compact) uses the 10 mean features; scenario B (full) uses all 30.

C. Preprocessing and Models

Features are standardised inside each training fold through a single pipeline to avoid leakage. The models are logistic regression (L2, $C=1$), an RBF-kernel SVM ($C=1$, probability estimates by Platt scaling), a random forest (500 trees) [11], and gradient boosting (100 trees, depth 3, learning rate 0.1) [12], plus a majority-class baseline. All use scikit-learn 1.8.0 [13] with default settings otherwise; no hyperparameter tuning was performed. The random seed is 42.

D. Evaluation Protocol

Data are split once into a stratified training set (455 samples) and a held-out test set (114 samples). Models are compared on the training set only, using repeated stratified 5-fold cross-

validation with 10 repeats (50 folds). The model with the highest mean cross-validated ROC-AUC in each scenario is refit on the full training set and evaluated once on the test set at a decision threshold of 0.5. We report accuracy, precision, recall, F1 and ROC-AUC, plus the Brier score [9], with 95% bootstrap confidence intervals (2,000 resamples) [10] for accuracy and AUC on the test set. Feature influence is measured by permutation importance on the test set (30 repeats, drop in ROC-AUC).

VII. RESULTS

A. Cross-Validated Model Comparison

Table 2 and Fig. 2 show the training-set cross-validation results. In both scenarios every model is far above the baseline. In scenario A the four models reach AUC 0.982–0.985, and logistic regression, SVM and gradient boosting are tied at 0.985 to three decimals; gradient boosting is selected by a margin below the third decimal, so the choice among them is not meaningful. In scenario B, logistic regression and the SVM are tied at 0.994 and the tree ensembles are slightly lower (0.988); the SVM is selected by a margin below the third decimal. Differences between models are smaller than the fold-to-fold standard deviations, which are 0.006–0.011 for AUC. Note that folds share training data, so these standard deviations are not standard errors of the mean.

B. Held-Out Test Performance

Table 3 gives the single test-set evaluation of the selected model in each scenario. In scenario A the gradient boosting model reaches accuracy 0.947 and AUC 0.979. In scenario B the SVM reaches accuracy 0.965 and AUC 0.995, with the Brier score halving from 0.046 to 0.022. Fig. 3 and Fig. 4 show the ROC curves and confusion matrices. In scenario A the model missed 4 of 42 malignant samples and flagged 2 of 72 benign samples as malignant; in scenario B it missed 2 of 42 and flagged 2 of 72. For a screening use case, a missed malignancy is usually the costlier error, so the lower miss count in scenario B is favourable, but a difference of two samples in a 114-sample test set is well within sampling variation. The 95% confidence intervals for accuracy (0.904–0.982 vs. 0.930–0.991) and for AUC (0.947–0.998 vs. 0.982–1.000) overlap substantially. The test AUC for both scenarios is consistent with the cross-validated values (0.985 and 0.994).

C. Feature Importance

Fig. 5 shows permutation importance. In scenario A, two features dominate: mean area (mean AUC drop 0.110) and mean concave points (0.084), followed by mean texture (0.024); all other features contribute less than 0.003. In scenario B no single feature exceeds a drop of 0.007 (worst texture, 0.006), and the standard deviations are of similar size as the means. This does not mean that the features are unimportant: 21 of the 435 feature pairs have an absolute correlation above 0.9 (maximum 0.998), so permuting one feature leaves correlated copies that carry the same information. The B-scenario ranking is therefore unstable and should not be over-interpreted [8].

VIII. DISCUSSION

The central finding is that a compact set of ten mean nucleus measurements already supports high-quality classifica-

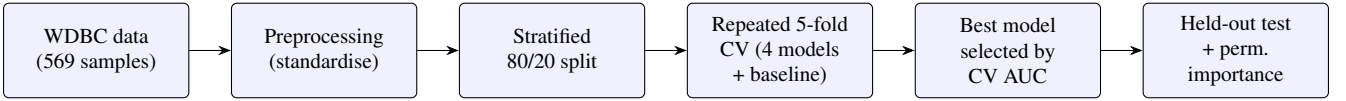


Figure 1: Experimental pipeline used in this study.

TABLE 2. CROSS-VALIDATION RESULTS ON THE TRAINING SET (MEAN $\pm$ SD OVER 50 FOLDS; MALIGNANT = POSITIVE CLASS; SELECTED MODEL IN BOLD)

Scenario	Model	Accuracy	F1	Recall	ROC-AUC
A: mean only	Majority baseline	0.626 $\pm$ 0.000	0.000 $\pm$ 0.000	0.000 $\pm$ 0.000	0.500 $\pm$ 0.000
	Logistic Regression	0.939 $\pm$ 0.020	0.916 $\pm$ 0.028	0.892 $\pm$ 0.043	0.985 $\pm$ 0.007
	SVM (RBF)	0.942 $\pm$ 0.021	0.918 $\pm$ 0.031	0.881 $\pm$ 0.053	0.985 $\pm$ 0.009
	Random Forest	0.937 $\pm$ 0.023	0.916 $\pm$ 0.031	0.911 $\pm$ 0.043	0.982 $\pm$ 0.011
	Gradient Boosting	0.940 $\pm$ 0.020	0.919 $\pm$ 0.027	0.916 $\pm$ 0.039	0.985 $\pm$ 0.008
B: all 30	Majority baseline	0.626 $\pm$ 0.000	0.000 $\pm$ 0.000	0.000 $\pm$ 0.000	0.500 $\pm$ 0.000
	Logistic Regression	0.972 $\pm$ 0.016	0.962 $\pm$ 0.022	0.947 $\pm$ 0.037	0.994 $\pm$ 0.007
	SVM (RBF)	0.970 $\pm$ 0.019	0.959 $\pm$ 0.025	0.948 $\pm$ 0.037	0.994 $\pm$ 0.006
	Random Forest	0.956 $\pm$ 0.024	0.940 $\pm$ 0.034	0.928 $\pm$ 0.051	0.988 $\pm$ 0.009
	Gradient Boosting	0.956 $\pm$ 0.025	0.939 $\pm$ 0.035	0.926 $\pm$ 0.049	0.988 $\pm$ 0.010

TABLE 3. HELD-OUT TEST PERFORMANCE ($n=114$)

Metric	A: Grad. Boosting	B: SVM (RBF)
Accuracy	0.947	0.965
95% CI	0.904–0.982	0.930–0.991
Precision	0.950	0.952
Recall	0.905	0.952
F1	0.927	0.952
ROC-AUC	0.979	0.995
95% CI	0.947–0.998	0.982–1.000
Brier score	0.046	0.022
Baseline accuracy	0.632	0.632

TABLE 4. THREATS TO VALIDITY

Threat	Explanation
Small sample	Only 569 samples (114 in the test set); intervals are wide and one split was used.
Single source	One public dataset from one source; results may not transfer to other laboratories or imaging setups.
Untuned models	Default hyperparameters were used; tuning might change the model ranking slightly.
Correlated features	21 feature pairs have $ r > 0.9$; permutation importance is unreliable for them.
Ceiling effect	AUC near 0.99 leaves little room to separate models or scenarios.
Fairness	The dataset contains no patient demographics, so subgroup performance could not be audited.
Clinical use	This is a retrospective benchmark; no prospective or clinical validation was performed.

tion on this dataset: cross-validated AUC is 0.985 in scenario A against 0.994 in scenario B. The gain from the full feature set is consistent in direction across logistic regression, SVM and the tree ensembles, but it is small—about one fold-to-fold standard deviation—and the held-out intervals overlap, so the evidence supports “at most a small improvement” rather than a confirmed one. The Brier score does favour scenario B (0.022 vs. 0.046), indicating better-calibrated probabilities, but it too comes from a single 114-sample test set.

Three cautions apply. First, scores near the ceiling leave little room to discriminate between models; the selection of gradient boosting in A and of the SVM in B reflects differences below the third decimal and should not be read as a recommendation. Second, a simple, interpretable model such as logistic regression performs as well as the more complex ensembles in both scenarios, which is consistent with arguments for preferring interpretable models in high-stakes settings [6]. Third, the descriptors in this dataset are computed from curated, digitised images collected at one source; performance on images from other laboratories, scanners or staining protocols is not shown by this study.

IX. THREATS TO VALIDITY

Table 4 lists the main limitations. In addition, the decision threshold of 0.5 is a default; a screening setting would normally choose a threshold that favours sensitivity, and this

would change precision and recall.

X. CONCLUSION

This paper compared four machine learning models and a baseline for classifying breast masses as malignant or benign on the Wisconsin Diagnostic Breast Cancer dataset. With only the 10 mean nucleus measurements, the selected gradient boosting model reached a held-out accuracy of 0.947 and AUC of 0.979; with all 30 measurements the selected SVM reached 0.965 and 0.995. The difference between the two scenarios is small relative to the uncertainty of the evaluation, and the choice among models is not meaningful at this performance level. The results show that standard models are accurate on this curated dataset, but they do not establish clinical usefulness. Future work will use repeated outer splits or nested cross-validation, hyperparameter tuning, threshold selection for a target sensitivity, external validation on other cytology datasets, and integration with explanation modules, followed by evaluation under clinical supervision.

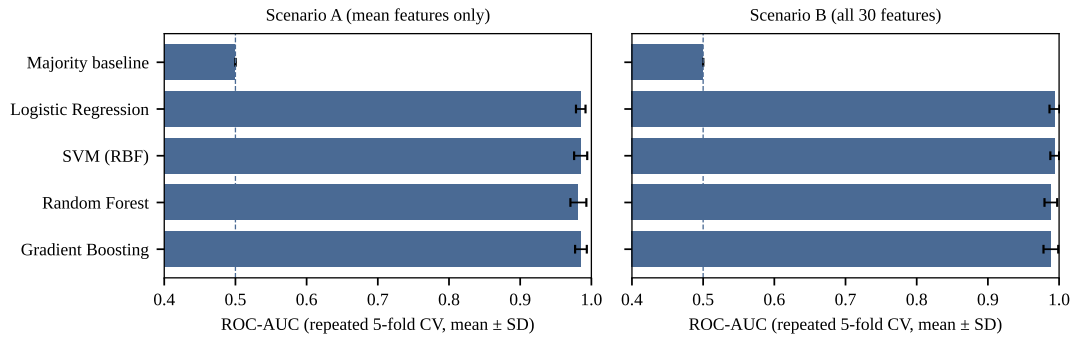


Figure 2: Cross-validated ROC-AUC by model and scenario. The dashed line marks chance level.

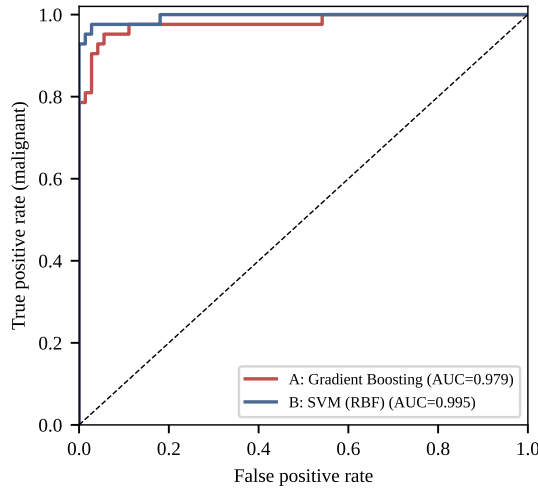


Figure 3: Test-set ROC curves for the selected model in each scenario.

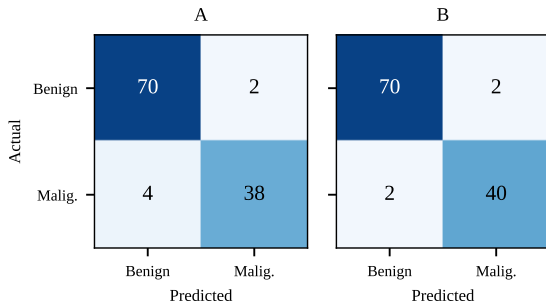


Figure 4: Test-set confusion matrices (A and B).

DATA AND CODE AVAILABILITY

The dataset is publicly available from the UCI Machine Learning Repository (Breast Cancer Wisconsin, Diagnostic) [4] and is also distributed with scikit-learn. All experiments use random seed 42 and scikit-learn 1.8.0.

ACKNOWLEDGMENT

The author would like to thank Sanjivani University and the faculty members for their guidance and support during this research.

REFERENCES

- [1] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," in *Proc. IS&T/SPIE Int. Symp. Electronic Imaging: Science and Technology*, vol. 1905, 1993, pp. 861–870.
- [2] W. H. Wolberg and O. L. Mangasarian, "Multisurface method

of pattern separation for medical diagnosis applied to breast cytology," *Proc. Natl. Acad. Sci. USA*, vol. 87, no. 23, pp. 9193–9196, 1990.

- [3] O. L. Mangasarian, W. N. Street, and W. H. Wolberg, "Breast cancer diagnosis and prognosis via linear programming," *Operations Research*, vol. 43, no. 4, pp. 570–577, 1995.
- [4] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Breast Cancer Wisconsin (Diagnostic)," UCI Machine Learning Repository, 1995, doi: 10.24432/C5DW2B.
- [5] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, 2015.
- [6] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, pp. 206–215, 2019.
- [7] G. C. Cawley and N. L. C. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *J. Machine Learning Research*, vol. 11, pp. 2079–2107, 2010.
- [8] 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, art. 82, 2021.
- [9] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, 1950.
- [10] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY, USA: Chapman & Hall, 1993.
- [11] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [12] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
- [13] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

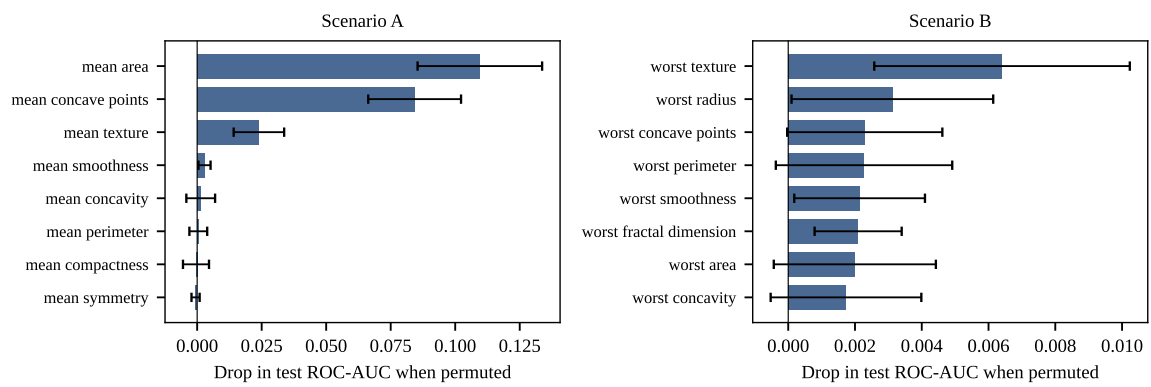


Figure 5: Permutation importance (test set) for the selected model in each scenario; top eight features, mean $\pm$ SD over 30 repeats.