Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Calibration and Threshold Selection of Machine Learning Classifiers for Breast Tumour Diagnosis: A Decision-Support Study for AI-Driven Clinical Screening

KARTIK PHOPASE

kartik.phopse_24uds@sanjivani.edu.in

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

Research Article

Keywords: Machine Learning, Probability Calibration, Breast Cancer Diagnosis, Clinical Decision Support, Threshold Selection, Healthcare Analytics

Posted Date: October 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11491081/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.

Calibration and Threshold Selection of Machine Learning Classifiers for Breast Tumour Diagnosis: A Decision-Support Study for AI-Driven Clinical Screening

Kartik Phopase
ORCID: 0009-0002-4168-179X
Department of Artificial Intelligence and Data Science
Sanjivani University, Maharashtra, India
Preprint, October 2026

Abstract—Machine learning classifiers for tumour diagnosis are usually judged by accuracy, but clinical decision support also needs trustworthy probabilities and a justified decision threshold. This paper studies these two aspects on the public Wisconsin Diagnostic Breast Cancer dataset (569 samples, 37.3% malignant) distributed with scikit-learn. Four 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 split. Two scenarios are examined: (A) the 10 mean-value features only and (B) all 30 features. Cross-validated ROC-AUC is 0.982–0.985 in scenario A and 0.988–0.994 in scenario B, and differences between models are not statistically significant under a corrected resampled t-test (all $p > 0.14$). The selected models reach held-out accuracy of 0.947 (A, gradient boosting) and 0.965 (B, SVM), with AUC 0.979 and 0.995. Logistic regression and SVM probabilities already have low calibration error; post-hoc sigmoid calibration worsened them, whereas isotonic calibration lowered the calibration error of random forest and gradient boosting, although Brier-score differences are small relative to bootstrap intervals. A threshold chosen on out-of-fold predictions to reach 95% sensitivity gave a held-out sensitivity of 0.929 in both scenarios, below the target, showing that threshold selection is itself uncertain at this sample size. These findings are relevant for designing the risk-estimation component of AI-driven clinical decision-support systems.

Index Terms—*Machine Learning, Probability Calibration, Breast Cancer Diagnosis, Clinical Decision Support, Threshold Selection, Healthcare Analytics*

I. INTRODUCTION

Artificial Intelligence (AI) is increasingly used to support clinical decisions such as screening, triage, and diagnosis. For breast cancer, fine needle aspiration (FNA) cytology provides a minimally invasive way to examine a breast mass, and computer-assisted analysis of digitised FNA images has been studied for decades [1], [2]. The resulting Wisconsin Diagnostic Breast Cancer (WDBC) dataset, available from the UCI Machine Learning Repository [3], has become a standard benchmark on which many classifiers reach very high accuracy.

High accuracy alone is not enough for a decision-support component. A system that passes a malignancy probability to a downstream triage rule needs that probability to be calibrated, and it needs a decision threshold that reflects the higher cost of missing a malignant tumour. Calibration is often neglected in the reporting of predictive models [4]. This paper therefore asks a focused empirical question: on a benchmark where accuracy is near its ceiling, how do standard classifiers differ in probability calibration and in the reliability of a sensitivity-constrained threshold? The main contributions are:

- A reproducible comparison of four machine learning models and a baseline on WDBC under one leakage-free protocol with repeated cross-validation, a corrected significance test, and bootstrap confidence intervals.
- Two feature-set scenarios—mean-value features only (A) and the full 30 features (B)—evaluated identically.
- An empirical look at post-hoc calibration (sigmoid and isotonic) and at how well a threshold tuned on out-of-fold predictions transfers to unseen data, reported with its uncertainty.

II. PROBLEM STATEMENT

A. Accuracy-Only Reporting

Many studies on WDBC report accuracy or AUC for several models and rank them, even when differences are smaller than the sampling noise. Without repeated resampling and an appropriate significance test, such rankings can be misleading.

B. Need for Calibrated Risk and a Justified Threshold

In screening, the two error types have unequal costs. A classifier is only useful as a decision aid if its probabilities mean what they say and if the chosen threshold achieves the intended sensitivity on new patients. It is unclear how well standard models satisfy the first requirement out of the box, and how stable the second is when the threshold is tuned on limited data.

III. RESEARCH OBJECTIVES

1. To compare machine learning models that classify a breast mass as malignant or benign, using a protocol that separates model comparison from final evaluation.

2. To evaluate probability calibration before and after post-hoc calibration, and to test how reliably a sensitivity-constrained threshold chosen on training data transfers to a held-out set, in a reduced (A) and a full (B) feature scenario.

IV. LITERATURE REVIEW

A. Machine Learning for Breast Tumour Diagnosis

Street et al. described the extraction of nuclear features from digitised FNA images for breast tumour diagnosis, which led to the WDBC features used here [1]. Earlier, Wolberg and Mangasarian applied a multisurface pattern-separation method to breast cytology data [2]. The dataset is distributed through the UCI repository [3] and is also bundled with scikit-learn.

B. Probability Calibration

Niculescu-Mizil and Caruana compared the probabilities produced by many learning algorithms and showed that methods such as boosted trees and SVMs tend to give distorted probabilities that Platt scaling or isotonic regression can improve [5]. Platt proposed fitting a sigmoid to SVM outputs [6], and Zadrozny and Elkan used isotonic regression to calibrate classifier scores [7]. Guo et al. showed that modern neural networks are often poorly calibrated and can be corrected by simple post-hoc scaling [8].

C. Thresholds, Costs and Evaluation Bias

Elkan showed that the optimal decision threshold depends on the relative costs of the two error types [9]. Vickers and Elkin proposed decision curve analysis to assess the clinical value of prediction models across threshold probabilities [10]. Cawley and Talbot showed that tuning and evaluating on the same data produces optimistically biased performance estimates [11], which motivates the strict separation of training-set comparison and held-out testing used here.

V. RESEARCH GAP

Benchmark studies on WDBC typically report accuracy and rank models. Fewer examine calibration alongside discrimination, and fewer still check whether a sensitivity target set on training data is met on unseen data. This study addresses that narrower gap under a single protocol with uncertainty estimates.

VI. RESEARCH METHODOLOGY

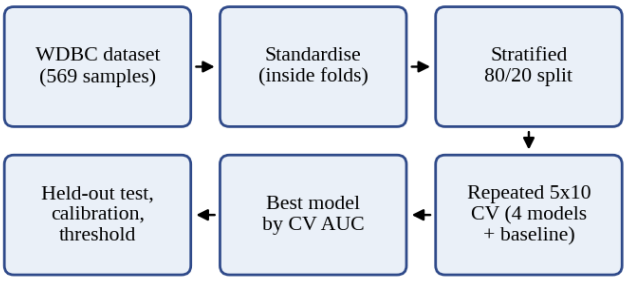

Fig. 1. Experimental pipeline used in this study.

A. Dataset

The data are the WDBC dataset [1], [3]: 569 samples, 30 real-valued attributes, and no missing values. Each attribute is derived from a digitised FNA image and describes the cell nuclei. Ten characteristics (radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension) are each summarised by their mean, standard error, and "worst" value (mean of the three largest values). Of the 569 samples, 212 (37.3%) are malignant and 357 are benign. Table I gives the breakdown.

TABLE I. INPUT VARIABLES BY FEATURE GROUP

Feature group	Count	Used in	Examples
Mean values	10	A and B	mean radius, mean area
Standard errors	10	B only	area error, texture error
Worst values	10	B only	worst radius, worst concavity

B. Target and Scenarios

The target is malignant = 1 (positive class) and benign = 0, because failing to detect a malignant tumour is the costlier error. The majority-class baseline predicts benign and reaches an accuracy of 0.632 on the test set. Scenario A uses only the 10 mean-value features. Scenario B uses all 30 features.

C. Preprocessing and Models

Numeric variables 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 [12] (C=1), a random forest (500 trees) [13], and gradient boosting (100 trees, depth 3, learning rate 0.1) [14], plus a majority-class baseline. All use scikit-learn 1.8.0 [15] with default settings otherwise; no hyperparameter tuning was performed, and the random seed is 42. The SVM obtains probabilities through scikit-learn's built-in internal Platt scaling, so its "raw" probabilities are already sigmoid-calibrated.

D. Evaluation Protocol

Data are split once into a stratified training set (455 samples) and a held-out test set (114 samples; 42 malignant, 72 benign). 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. Differences between the selected model and each alternative are tested with the corrected resampled t-test [18] on per-fold AUC, using the variance correction factor (1/k + 1/4) with k=50 folds, where 1/4 is the ratio of test-fold to training-fold size. We report accuracy, precision, recall (sensitivity), specificity, F1, and ROC-AUC [20], the Brier score [16], and the expected calibration error (ECE, 10 equal-width bins) [17], with 95% percentile-bootstrap confidence intervals [19] (2,000 resamples) on the test set.

Calibration. Sigmoid and isotonic calibration are fitted with 5-fold internal cross-validation on the training set only and evaluated on the test set (scenario B, all four models). *Threshold.* Out-of-fold (OOF) probabilities are obtained on the training set with 10-fold cross-validation, and the threshold is set to the largest value whose OOF sensitivity is at least 0.95, which maximises specificity under that constraint. It is applied once to the test set and compared with the default threshold of 0.5. *Importance.* 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 II. CROSS-VALIDATION RESULTS ON THE TRAINING SET (MEAN ± SD OVER 50 FOLDS; MALIGNANT = POSITIVE CLASS)

Scenario	Model	Accuracy	F1	Recall	ROC-AUC
A: mean	Majority base.	0.626±0.000	0.000±0.000	0.000±0.000	0.500±0.000
	Logistic Reg.	0.939±0.020	0.916±0.028	0.892±0.043	0.985±0.007
	SVM (RBF)	0.942±0.021	0.918±0.031	0.881±0.053	0.985±0.009
	Random Forest	0.937±0.023	0.916±0.031	0.911±0.043	0.982±0.011
	Grad. Boosting	0.940±0.020	0.919±0.027	0.916±0.040	0.985±0.008
B: all 30	Majority base.	0.626±0.000	0.000±0.000	0.000±0.000	0.500±0.000
	Logistic Reg.	0.972±0.016	0.962±0.023	0.947±0.037	0.994±0.007
	SVM (RBF)	0.970±0.019	0.959±0.026	0.948±0.037	0.994±0.006
	Random Forest	0.956±0.025	0.940±0.034	0.928±0.051	0.988±0.009
	Grad. Boosting	0.956±0.026	0.939±0.036	0.926±0.050	0.988±0.010

Table II and Fig. 2 show the training-set cross-validation results. In scenario A all four models are close (AUC 0.982–0.985, accuracy 0.937–0.942), and differences between models are far smaller than the fold-to-fold standard deviations. Gradient boosting has the highest mean AUC by a margin of 0.0003 and is selected, but the corrected t-test finds no significant difference from logistic regression ($p = 0.92$), the SVM ($p = 0.89$) or random forest ($p = 0.32$). In scenario B the SVM and logistic regression (AUC 0.994) are about 0.005 above random forest and gradient boosting (0.988–0.989); the SVM is selected by a margin of 0.0003 over logistic regression, and no difference from any other model is significant ($p = 0.88$, 0.15, 0.15 for logistic regression, random forest, gradient boosting). The choice among the top models is therefore not meaningful. Moving from A to B raises mean AUC for every model and mean accuracy by 0.016–0.034.

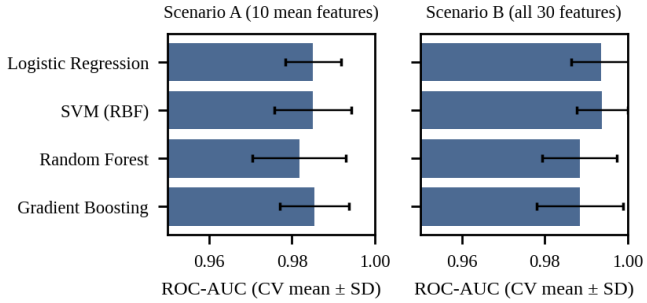

Fig. 2. Cross-validated ROC-AUC by model and scenario (mean ± SD; axis truncated at 0.95). The majority baseline has AUC 0.5 (Table II).

B. Held-Out Test Performance

TABLE III. HELD-OUT TEST PERFORMANCE (n=114; THRESHOLD 0.5)

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 (sensitivity)	0.905	0.952
Specificity	0.972	0.972
F1	0.927	0.952
ROC-AUC	0.979	0.995
95% CI	0.946–0.998	0.983–1.000
Brier score	0.046	0.022
ECE	0.044	0.020
Baseline accuracy	0.632	0.632

Table III gives the single test-set evaluation of the selected model in each scenario. Scenario A reaches accuracy 0.947 and AUC 0.979; scenario B reaches accuracy 0.965 and AUC 0.995, far above the baseline. The intervals overlap (AUC 0.946–0.998 vs. 0.983–1.000), so the test set alone cannot establish that B is better than A; the cross-validation results (Section VII-A) point in the same direction, but the gain is modest. Fig. 3 and Fig. 4 show the ROC curves and confusion matrices. In scenario A the model misses 4 of 42 malignant tumours and flags 2 of 72 benign cases; in scenario B it misses 2 and flags 2.

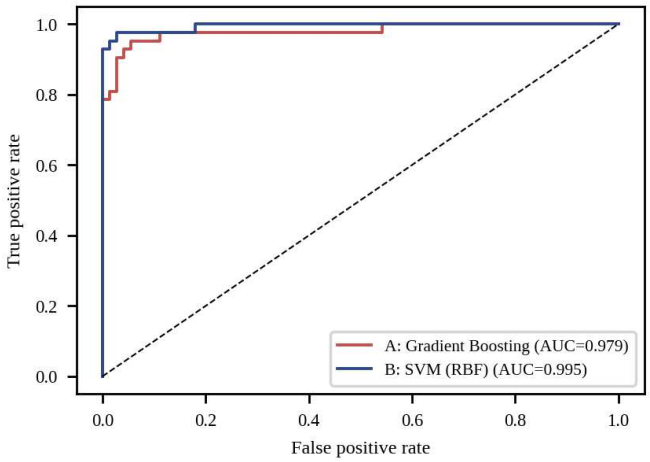

Fig. 3. Test-set ROC curves for the selected model in each scenario.

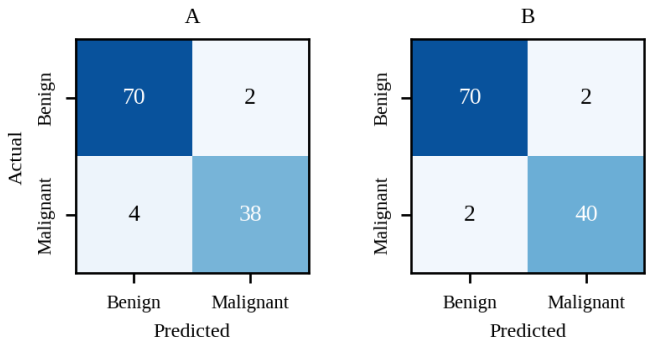

Fig. 4. Test-set confusion matrices (A and B; threshold 0.5).

C. Probability Calibration

TABLE IV. TEST-SET BRIER SCORE / ECE BEFORE AND AFTER POST-HOC CALIBRATION (SCENARIO B)

Model	Raw	Sigmoid	Isotonic
Logistic Reg.	0.021 / 0.030	0.031 / 0.070	0.017 / 0.021
SVM (RBF)	0.022 / 0.020	0.026 / 0.048	0.021 / 0.028
Random Forest	0.031 / 0.059	0.029 / 0.034	0.029 / 0.032
Grad. Boosting	0.033 / 0.038	0.026 / 0.051	0.024 / 0.028

Table IV and Fig. 6 compare calibration. Before any post-hoc step, logistic regression and the SVM have the lowest ECE (0.030 and 0.020), while random forest has the highest (0.059). Isotonic calibration lowered ECE for logistic regression (0.030 to 0.021), random forest (0.059 to 0.032) and gradient boosting (0.038 to 0.028) and slightly raised it for the SVM (0.020 to 0.028). Sigmoid calibration made logistic regression clearly worse (Brier 0.021 to 0.031, ECE 0.030 to 0.070) and also raised the SVM's and gradient boosting's ECE. A plausible reason for the SVM is that its probabilities were already Platt-scaled, so a second sigmoid adds noise. All Brier differences between variants are at most 0.013, small relative to the bootstrap intervals (for example 0.006–0.042 for raw logistic regression), and ECE from 114 samples is itself noisy. AUC is essentially unchanged by calibration (0.989–0.997 across all variants).

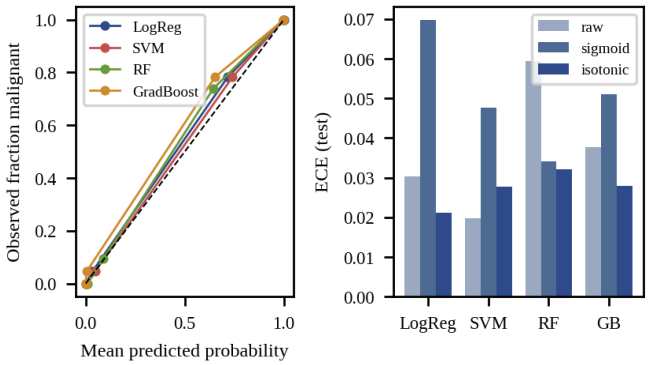

Fig. 6. Left: reliability curves of the raw test-set probabilities (scenario B; 5 equal-count bins). Right: ECE for raw, sigmoid and isotonic probabilities.

D. Sensitivity-Constrained Threshold

TABLE V. TEST-SET RESULTS WITH THE DEFAULT AND THE OOF-TUNED THRESHOLD (42 MALIGNANT, 72 BENIGN)

Setting	Thr.	Sens.	Spec.	FN	FP
A, default	0.500	0.905	0.972	4	2
A, OOF-tuned	0.322	0.929	0.944	3	4
B, default	0.500	0.952	0.972	2	2
B, OOF-tuned	0.556	0.929	0.986	3	1

The OOF-tuned thresholds are 0.322 (A) and 0.556 (B), and on the training OOF predictions they give a sensitivity of 0.953. On the test set, however, sensitivity is 0.929 in both scenarios (39 of 42; exact 95% interval 0.81–0.99), below the 0.95 target. In scenario A the lower threshold recovered one extra malignant case (4 to 3 misses) at the cost of two more false alarms; in scenario B the tuned threshold was higher than 0.5 (because OOF sensitivity already exceeded 0.95 at 0.5) and cost one additional miss (2 to 3). With 42 malignant test cases, one patient changes sensitivity by 2.4 percentage points, so these differences are within sampling noise; the main message is that a threshold fixed on training data does not guarantee the target sensitivity on new data.

E. Feature Importance

Fig. 5 shows permutation importance. In scenario A the largest drops in AUC are for mean area (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 0.006 (worst texture), and the standard deviations are of similar size, so the ranking is unstable. This is expected when many highly correlated features (radius, perimeter, area and their worst values) carry overlapping information: permuting one feature leaves its correlated partners to compensate. The result should not be read as showing that these features are unimportant.

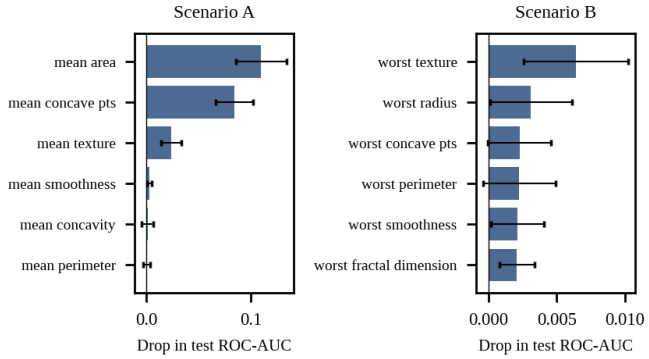

Fig. 5. Permutation importance (test set) for the selected model in each scenario; top six features, mean ± SD over 30 repeats.

VIII. DISCUSSION

Four points stand out. First, on this benchmark discrimination is near its ceiling: every model reaches a cross-validated AUC above 0.98, and no model is significantly better than another. Model choice therefore matters less than the evaluation design, and rankings based on a single split would be unreliable. Second, the full 30-feature set (B) improves cross-validated AUC and accuracy for all four models relative to the mean-only set (A), although the held-out intervals overlap. Third, calibration is not automatic and not always improved by post-hoc methods: logistic regression and the SVM were already well calibrated, an extra sigmoid layer made them worse, and tree ensembles benefited most from isotonic calibration. A practical rule is to check calibration on held-out data before and after any recalibration rather than to apply it by default. Fourth, threshold selection carries its own uncertainty: a sensitivity target met on out-of-fold training predictions was not met on the test set.

For AI-driven clinical decision support, these results suggest that the risk-estimation component should report calibrated probabilities with uncertainty, and that thresholds should be set with an explicit margin or a cost model [9], [10] and validated on independent data. A more conservative sensitivity target, or a lower confidence bound on sensitivity, may be preferable to a point target; this was not tested here. Two cautions apply. The WDBC features are computed from expert-selected FNA images, so the results do not describe end-to-end screening performance, and this study is a methodological benchmark, not a clinical validation.

IX. THREATS TO VALIDITY

TABLE VI. THREATS TO VALIDITY

Threat	Explanation
Small sample	569 samples; 114 in the test set and only 42 malignant, so intervals are wide and one split was used.
Single dataset	One data source; near-ceiling accuracy may not transfer to other populations, scanners or labs. No external validation.
Untuned models	Default hyperparameters were used; tuning might change the model ranking, although differences are already small.
Arbitrary selection	Selection by mean CV AUC rests on margins of 0.0003, so the selected model is effectively arbitrary.
Noisy calibration metrics	ECE with 10 bins on 114 samples is noisy; Brier differences lie inside bootstrap intervals.
Correlated features	Permutation importance is unreliable when features are strongly correlated.
Subgroups and fairness	The dataset has no demographic variables, so subgroup performance and fairness could not be assessed.

In addition, the 0.95 sensitivity target and the “largest threshold” rule are modelling choices; other targets or cost ratios could give different conclusions.

X. CONCLUSION

This paper compared four machine learning models and a baseline for classifying breast masses on the WDBC dataset, with attention to calibration and thresholds. All models were strongly discriminative (cross-validated AUC 0.982–0.994) and statistically indistinguishable. The selected models reached held-out accuracy of 0.947 (A, gradient boosting) and 0.965 (B, SVM) with AUC 0.979 and 0.995. Logistic regression and the SVM were already well calibrated, post-hoc sigmoid calibration was harmful for them, and isotonic calibration helped random forest and gradient boosting. A sensitivity-constrained threshold set on training data did not reach its target on held-out data. For AI-based clinical decision-support systems, the risk-estimation component should therefore be evaluated for calibration and threshold stability, not only for accuracy.

Future work will use larger and multi-centre datasets, nested cross-validation, hyperparameter tuning, decision-curve analysis, cost-based thresholds, subgroup and fairness audits, and external validation.

DATA AND CODE AVAILABILITY

The dataset is publicly available from the UCI Machine Learning Repository (Breast Cancer Wisconsin, Diagnostic) [3] and is bundled with scikit-learn. All experiments use random seed 42 and scikit-learn 1.8.0. The complete analysis script accompanies this preprint as a supplementary file.

ETHICS STATEMENT

The study uses a public, de-identified benchmark dataset and no new patient data. The models are not intended for clinical use.

ACKNOWLEDGMENT

The author would like to thank Sanjivani University and the faculty members for their guidance and support during this research. An AI assistant (Claude, Anthropic) was used to help prepare the analysis code, figures and manuscript draft; the author is responsible for the content.

REFERENCES

[1] 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.

[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] D. Dua and C. Graff, “UCI Machine Learning Repository,” Univ. of California, Irvine, School of Information and Computer Science, 2019. [Online]. Available: https://archive.ics.uci.edu

[4] 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.

[5] A. Niculescu-Mizil and R. Caruana, “Predicting good probabilities with supervised learning,” in *Proc. 22nd Int. Conf. on Machine Learning (ICML)*, 2005, pp. 625–632.

[6] J. C. Platt, “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,” in *Advances in Large Margin Classifiers*, MIT Press, 1999, pp. 61–74.

[7] B. Zadrozny and C. Elkan, “Transforming classifier scores into accurate multiclass probability estimates,” in *Proc. 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2002, pp. 694–699.

[8] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in *Proc. 34th Int. Conf. on Machine Learning (ICML)*, PMLR 70, 2017, pp. 1321–1330.

[9] C. Elkan, “The foundations of cost-sensitive learning,” in *Proc. 17th Int. Joint Conf. on Artificial Intelligence (IJCAI)*, 2001, pp. 973–978.

[10] A. J. Vickers and E. B. Elkin, “Decision curve analysis: a novel method for evaluating prediction models,” *Medical Decision Making*, vol. 26, no. 6, pp. 565–574, 2006.

[11] 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.

[12] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

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

[14] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[15] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *J. Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[16] G. W. Brier, “Verification of forecasts expressed in terms of probability,” *Monthly Weather Review*, vol. 78, no. 1, pp. 1–3, 1950.

[17] M. P. Naeini, G. F. Cooper, and M. Hauskrecht, “Obtaining well calibrated probabilities using Bayesian binning,” in *Proc. 29th AAAI Conf. on Artificial Intelligence*, 2015, pp. 2901–2907.

[18] C. Nadeau and Y. Bengio, “Inference for the generalization error,” *Machine Learning*, vol. 52, no. 3, pp. 239–281, 2003.

[19] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY, USA: Chapman & Hall, 1993.

[20] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.