

CALIBRATED RECOGNITION OF MALIGNANT NUCLEAR MORPHOLOGY FROM IMAGE-DERIVED FINE-NEEDLE ASPIRATE FEATURES

Dr. Vijay R.Rathod✉

Head, Department of Electronics and Telecommunication Engineering
ST Xaviers Technical Institute Mahim Mumbai 400016

Received: 20/01/2023

Revised: 04/02/2024

Accepted: 10/03/2024

ABSTRACT

Image-recognition studies in medicine frequently report high accuracy without separating model selection from performance estimation. This study evaluates whether compact machine-learning models can recognize malignant nuclear morphology from measurements extracted from digitized fine-needle aspirate images while controlling optimism. The public Wisconsin Diagnostic Breast Cancer dataset (569 observations; 212 malignant and 357 benign) was analysed using logistic regression, radial-basis-function support vector machines (RBF-SVM), and random forests. Hyperparameters were selected within four-fold inner cross-validation and evaluated by five-fold outer cross-validation repeated three times. Out-of-fold probabilities were averaged per observation. The RBF-SVM achieved an ROC area under the curve (AUC) of 0.995 (95% bootstrap confidence interval: 0.989–0.999), accuracy of 97.9%, sensitivity of 96.7%, specificity of 98.6%, and Brier score of 0.018. Logistic regression was numerically close, whereas random forest performance was lower. A feature-family ablation showed that the ten “worst” nuclear measurements retained an AUC of 0.992, while standard-error features alone fell to 0.942. These results support morphology-based computer assistance, but not autonomous diagnosis: the single-centre, image-derived feature set lacks external, temporal, and scanner validation.

KEYWORDS: Biomedical Image Recognition; Breast Cytology; Nested Cross-Validation; Calibration; Support Vector Machine

DOI: <https://doi.org/10.64149/gjaets.11.3.1-5>

1. INTRODUCTION

Digitized cytology converts visually assessed nuclear shape and texture into reproducible quantitative measurements. In fine-needle aspirate (FNA) analysis, cell-nucleus boundaries can be represented by radius, perimeter, area, smoothness, compactness, concavity, symmetry, texture, and fractal dimension. The Wisconsin Diagnostic Breast Cancer (WDBC) resource was created from such image-processed measurements and remains a useful controlled benchmark for studying the behaviour of diagnostic classifiers [1–3].

The central methodological problem is not whether a flexible algorithm can fit this benchmark, but whether its reported performance is protected from tuning bias and described beyond accuracy. Clinical prediction requires discrimination, class-specific error rates, and probability calibration [4,5]. When the same cross-validation loop selects hyperparameters and estimates performance, results can be optimistic [6]. This issue is pronounced in small datasets with correlated predictors.

Accordingly, this study asks three questions: (i) which of three compact classifiers provides the strongest nested-cross-validated discrimination; (ii) whether predicted probabilities remain reasonably calibrated; and (iii) how much information is retained by the mean, standard-error, and worst-value feature families. The contribution is an auditable evaluation design rather than a claim of clinical deployment.

2. BACKGROUND AND RELATED WORK

Street, Wolberg and Mangasarian developed an image-processing pipeline in which active contours delineated cell nuclei and geometric or intensity characteristics were computed from the segmented boundaries [1]. Subsequent work showed that linear-programming and machine-learning classifiers could distinguish benign and malignant aspirates using these image-derived descriptors [2]. The WDBC



release contains no missing values and provides 30 predictors: the mean, standard error (SE), and largest or “worst” value of ten base nuclear characteristics [3].

Support vector machines construct a maximum-margin boundary and can model nonlinear separation through a kernel [7]. Logistic regression offers a transparent linear probability model, while random forests combine randomized decision trees and naturally represent interactions [8]. On small biomedical datasets, however, headline accuracy can conceal two problems. First, a benign-dominated sample permits high accuracy despite missed malignancies. Second, an AUC near one does not guarantee reliable probabilities. Evaluation therefore needs sensitivity, specificity, F1, AUC, and a proper scoring rule such as the Brier score [4,5,9].

3. DATA AND EXPERIMENTAL DESIGN

The analysis used the WDBC copy distributed with scikit-learn, which reproduces the UCI dataset [3,10]. The 569 records comprise 212 malignant (37.3%) and 357 benign (62.7%) masses. Malignancy was defined as the positive class. The 30 continuous predictors were retained without univariate screening. This is an image-derived feature-recognition study: raw cytology images were not available, and conclusions therefore concern classification of extracted morphology rather than end-to-end image segmentation.

All scaling and fitting occurred inside resampling folds. Logistic regression and RBF-SVM were preceded by z-score standardization; random forests used the original feature scale. Class weights were balanced. Candidate grids were $C=\{0.05, 0.2, 1, 5\}$ for logistic regression; $C=\{0.5, 2, 8\}$ and $\gamma=\{\text{scale}, 0.01, 0.05\}$ for RBF-SVM; and combinations of maximum depth {unrestricted, 8} and minimum leaf size {1, 4} for a 160-tree forest. The random seed was fixed before analysis.

Component	Specification	Purpose
Outer validation	5 folds $\times$ 3 repeats	Low-bias internal assessment
Inner selection	4 stratified folds	Hyperparameter choice
Uncertainty	600 subject-level bootstraps	95% intervals
Primary outcome	ROC AUC	Threshold-independent discrimination
Secondary outcomes	Accuracy, sensitivity, specificity, F1, Brier	Error profile and calibration

4. METHODS

4.1 Repeated nested cross-validation

For each of 15 outer splits, grid search was performed only on the outer-training partition using four stratified inner folds and ROC AUC as the selection criterion. The selected estimator then generated probabilities for the untouched outer-test partition. Because the outer five-fold process was repeated three times, each observation received three independent out-of-fold probability estimates; these were averaged to produce one probability per mass. This preserves subject-level independence during scoring while reducing split-specific variance [6].

4.2 Metrics and uncertainty

Predictions were classified at a prespecified probability threshold of 0.50. Sensitivity was $TP/(TP+FN)$ for malignant masses; specificity was $TN/(TN+FP)$ for benign masses. AUC quantified ranking, and the Brier score measured mean squared probability error. Nonparametric 95% confidence intervals were generated by resampling observations with replacement 600 times. Calibration curves used eight quantile-based bins; they are descriptive because the sample is too small for stable subgroup calibration.

4.3 Feature-family ablation

The most frequently selected RBF-SVM setting ($C=8$, $\gamma=0.01$) was refitted in repeated five-fold cross-validation using four predictor sets: all 30 variables, the ten means, the ten SE values, and the ten worst values. This secondary analysis was deliberately fixed-parameter; it estimates information loss across the pre-defined families and is not directly interchangeable with the nested model-comparison estimate.



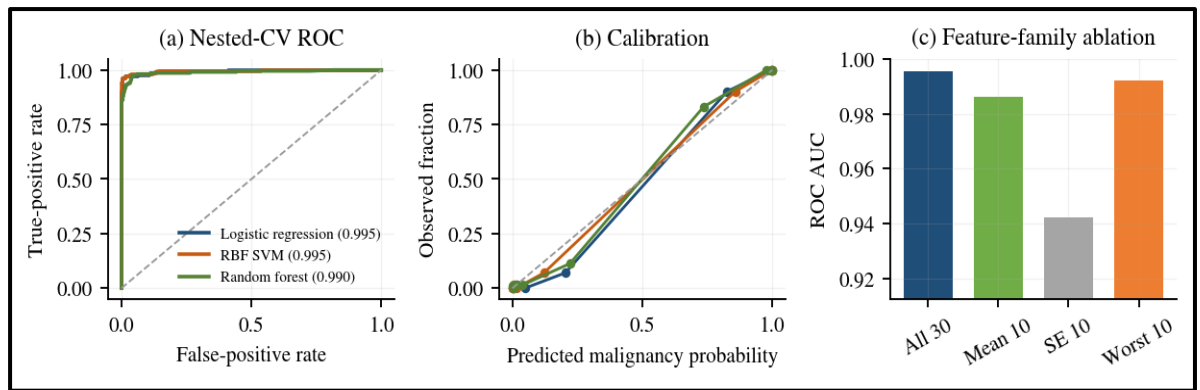


Figure 1. Discrimination, probability calibration, and pre-defined feature-family ablation.

5. Results

5.1 Comparative performance

The RBF-SVM ranked first by AUC (0.995) and Brier score (0.018). Its 97.9% accuracy corresponded to 205 true-positive malignant cases, seven false negatives, 352 true-negative benign cases, and five false positives. Logistic regression was nearly equivalent (AUC 0.995; accuracy 97.5%), indicating that most separation was already accessible through a regularized linear boundary.

Table 1. Repeated nested-cross-validated performance (malignant = positive class).

Model	Acc. %	Sens. %	Spec. %	AUC	Brier	F1
RBF SVM	97.9	96.7	98.6	0.995	0.018	0.972
Logistic regression	97.5	97.2	97.8	0.995	0.023	0.967
Random forest	96.0	93.9	97.2	0.990	0.031	0.945

The forest produced the lowest AUC (0.990) and the largest Brier score (0.031). Overlapping confidence intervals caution against interpreting the small RBF-SVM advantage as definitive.

5.2 Feature-family ablation

Table 2. Fixed-parameter RBF-SVM ablation across pre-defined nuclear feature families.

Feature family	Acc. %	Sens. %	Spec. %	AUC	Brier
All 30	98.6	97.2	99.4	0.996	0.017
Mean 10	94.6	90.6	96.9	0.986	0.040
SE 10	86.3	76.9	91.9	0.942	0.092
Worst 10	97.7	95.8	98.9	0.992	0.020

Worst-value features alone retained strong discrimination (AUC 0.992; 97.7% accuracy). Mean features were less sensitive to malignancy (90.6%), and SE features alone produced the largest deterioration (AUC 0.942; sensitivity 76.9%). This pattern suggests that extreme nuclear morphology contains information not fully represented by average or within-image variability summaries.

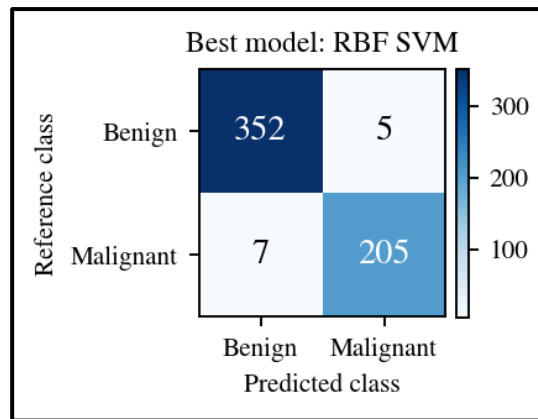


Figure 2. Subject-level confusion matrix for the nested-cross-validated RBF-SVM at threshold 0.50.

6. DISCUSSION

The study produced two substantive findings. First, both standardized logistic regression and RBF-SVM recognized malignant image-derived morphology with high discrimination. The close performance of the linear and nonlinear models matters: on this benchmark, algorithmic complexity contributed less than disciplined scaling, stratification, and separation of tuning from evaluation. A small dataset does not automatically justify a high-capacity image model when the input has already been condensed into biologically interpretable measurements.

Second, the worst-value family was substantially more informative than the SE family. This is plausible because a small group of markedly irregular nuclei may influence diagnosis even when average morphology remains closer to benign tissue. The result should not be interpreted causally, and correlated radius, perimeter, and area measures prevent attributing performance to one nuclear mechanism. Nevertheless, the ablation offers a practical hypothesis for parsimonious acquisition: retaining extreme morphology may be more valuable than retaining variability summaries alone.

6.1 Clinical and methodological implications

For computer-assisted cytology, the observed false-negative rate is more consequential than the overall accuracy. Seven malignant masses received an averaged probability below 0.50. Thresholds should therefore be selected from an explicit clinical utility function rather than inherited from a software default. Calibration also requires external assessment because probability reliability can shift when prevalence, staining, optics, segmentation, or patient spectrum changes [4,5]. A high internal AUC is evidence of separability within this dataset, not transportability.

6.2 Limitations

The dataset is small, historical, and derived from one diagnostic workflow. No raw images were re-segmented, so errors introduced by contour initialization or feature extraction could not be studied. Repeated cross-validation reuses observations across repeats and therefore narrows split variability without creating new external evidence. The bootstrap intervals quantify sampling uncertainty conditional on this dataset and pipeline; they do not include site or scanner heterogeneity. No subgroup variables were available for fairness assessment. Finally, model probabilities were generated by the SVM's internal calibration procedure and require independent recalibration before any prospective use.

Future work should validate the locked pipeline on temporally and geographically independent FNA cohorts, report decision-curve consequences, examine threshold-specific false negatives, and compare image-derived features with end-to-end segmentation under patient-level splits. An external validation protocol should be registered before inspecting outcome labels.

7. CONCLUSIONS

A repeated nested-cross-validation design showed that malignant nuclear morphology can be recognized accurately from the 30 WDBC image-derived features. The RBF-SVM achieved an AUC of 0.995 and 97.9% accuracy, but logistic regression was nearly indistinguishable. Worst-value nuclear measurements preserved most discrimination, whereas SE features alone were insufficiently sensitive. The results justify external validation of a compact decision-support model; they do not justify clinical automation.

8. DECLARATIONS

Data availability: The WDBC dataset is publicly available from the UCI Machine Learning Repository under CC BY 4.0 [3]. Code availability: The deterministic analysis script and derived tables should be deposited with the final submission. Ethics: No participants were recruited and no identifiable images were accessed; this secondary methodological analysis used public, de-identified image-derived features. Funding and conflicts of interest: [Authors to complete before submission].

REFERENCES

1. Street WN, Wolberg WH, Mangasarian OL. Nuclear feature extraction for breast tumor diagnosis. Proc SPIE 1905, Biomedical Image Processing and Biomedical Visualization. 1993;861–870. doi:10.1117/12.148698.
2. Wolberg WH, Street WN, Mangasarian OL. Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. Cancer Letters. 1994;77(2–3):163–171.
3. Wolberg W, Mangasarian O, Street N, Street W. Breast Cancer Wisconsin (Diagnostic) [dataset]. UCI Machine Learning Repository. 1993. doi:10.24432/C5DW2B.
4. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128–138. doi:10.1097/EDE.0b013e3181c30fb2.
5. Brier GW. Verification of forecasts expressed in terms of probability. Monthly Weather Review. 1950;78(1):1–3.
6. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006;7:91. doi:10.1186/1471-2105-7-91.
7. Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995;20:273–297. doi:10.1007/BF00994018.
8. Breiman L. Random forests. Machine Learning. 2001;45:5–32. doi:10.1023/A:1010933404324.
9. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006;27(8):861–874. doi:10.1016/j.patrec.2005.10.010.
10. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. Journal of Machine Learning Research. 2011;12:2825–2830.

CITE AN ARTICLE

Rathod V. R., (2024) Calibrated Recognition of Malignant Nuclear Morphology From Image-Derived Fine-Needle Aspirate Features. *Global Journal of Advanced Engineering and Technology Studies*, 11(3), 1-5 <https://doi.org/10.64149/gjaets.11.3.1-5>

