Proceedings of ICBioMed 2026 Symposium: Biomedical Engineering and Smart Medicine
DOI: 10.54254/2753-8818/2026.HZ38953

BI-RADS-Derived Breast Cancer Risk Classification from Mammograms

Zihan Wu

Marriotts Ridge High School, Marriottsville, USA
Zihanwu0123@gmail.com

Abstract. Mammography is a central imaging modality for breast-cancer screening, but interpretation remains difficult because suspicious findings can be subtle and breast tissue can be heterogeneous or dense. This study evaluates whether interpretable, hand-crafted digital image-processing features can support BI-RADS-derived low-risk versus high-risk classification using the INbreast full-field digital mammography dataset, which contained 410 mammograms from 108 anonymized patients. Labels were derived from BI-RADS categories (1-3: low risk; 4-6: high risk), and BI-RADS 0 examinations were excluded from the assigned modeling sets. A leakage-aware patient-level split produced 240 training, 84 validation, and 86 test images. The image-processing pipeline explored gray-level enhancement, convolution and filtering, binary-mask morphology, segmentation and region-of-interest construction, and then extracted 60 interpretable features spanning intensity statistics, histogram distribution, gray-level co-occurrence matrix texture, histogram of oriented gradients, and shape. Classical classifiers were compared using raw and train-fitted min-max-normalized features, with feature subsets selected using t-test-based incremental feature selection (tIFS) and Random-Forest recursive feature elimination (RF-RFE). On the held-out test set, the highest AUROC among evaluated final configurations was 0.7382 for RF-RFE Random Forest with min-max normalization; however, sensitivity was only 0.0417, demonstrating that overall discrimination or accuracy alone can obscure poor detection of the minority high-risk class. A more balanced RF-RFE SVM achieved AUROC 0.6969, sensitivity 0.5417, specificity 0.7581, and MCC 0.2869. These findings support interpretable image features as a transparent baseline while showing limitations from class imbalance, small sample size, and feature representation. The study represents BI-RADS-derived risk classification rather than pathology-confirmed cancer diagnosis.

Keywords: mammography, breast cancer, BI-RADS, INbreast, machine learning

1. Introduction

Breast cancer screening relies heavily on mammography, where radiologists assess findings such as masses, calcifications, architectural distortion, and asymmetry. The American College of Radiology Breast Imaging Reporting and Data System (BI-RADS) standardizes breast-imaging assessment and management categories [1]. In the INbreast dataset, BI-RADS 1 indicates no findings, BI-RADS 2 benign findings, BI-RADS 3 probably benign findings, BI-RADS 4 suspicious findings, BI-RADS 5

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.HZ38953

a high probability of malignancy, and BI-RADS 6 known cancer [2]. This study therefore uses BI-RADS to define an image-derived binary risk task rather than a pathology-confirmed benign-versus-malignant endpoint.

Mammographic interpretation remains difficult because lesions may be subtle or obscured by overlapping tissue. Computational approaches range from conventional computer-aided detection using engineered descriptors to deep-learning systems that learn representations directly from images. Deep learning has shown strong potential in mammography, but reproducibility, generalizability, external validation, and explainability remain important concerns [3-5]. Interpretable radiomic-style descriptors offer a complementary baseline because intensity, texture, edge, and shape measurements can be inspected directly and compared across models.

This study asks whether transparent image-derived features contain reproducible information associated with BI-RADS-derived risk. Low risk was defined as BI-RADS 1-3 and high risk as BI-RADS 4-6, with BI-RADS 0 excluded from the assigned modeling sets. We hypothesized that hand-crafted features would provide above-chance discrimination, while performance would depend on class imbalance, feature representation, preprocessing, and the sensitivity-specificity trade-off. The workflow was designed to keep training, validation, and held-out testing roles separate so that feature selection and model configuration did not use test performance.

2. Material and methods

2.1. Dataset, labels, and patient-level partitioning

The study used the publicly available INbreast full-field digital mammography resource [2]. The dataset contained 410 high-resolution grayscale mammograms from 108 anonymized patients, together with image-level metadata and annotations. BI-RADS 1-3 were assigned low_risk (0), BI-RADS 4-6 high_risk (1), and BI-RADS 0 examinations were excluded from the assigned training, validation, and test sets. All mammograms from the same patient were assigned to the same partition to reduce leakage across views or examinations. A random seed of 42 was used for classifiers exposing a random_state parameter. The composition of the final modeling dataset is summarized in Table 1.

Table 1. Composition of the modeling dataset. Each final feature matrix contained 60 image-derived features

Split	Images	Low risk	High risk	High-risk fraction
Training	240	184	56	0.2333
Validation	84	64	20	0.2381
Testing	86	62	24	0.2791

2.2. Image processing and final feature-generation pipeline

Exploratory image-processing demonstrations included gray-level enhancement, mean convolution, Gaussian smoothing, Sobel filtering, lesion-mask morphology, Canny edge detection, and ROI construction. These procedures were used to understand image-processing behavior and to perform visual quality checks, but they did not directly generate the final 60-feature modeling matrices. The final feature-generation pipeline instead created an Otsu-thresholded breast-tissue foreground mask

DOI: 10.54254/2753-8818/2026.HZ38953

for each indexed mammogram [6]. Expert lesion masks and annotation-derived bounding boxes were used for annotation checks and exploratory ROI demonstrations only; they were not used to compute the final five feature families.

From the Otsu-derived breast-tissue foreground mask and corresponding image region, 60 interpretable features were computed: intensity statistics describing brightness and variability; 10-bin histogram features describing coarse intensity distribution; GLCM texture descriptors including contrast, homogeneity, energy, and correlation across 0°, 45°, 90°, and 135° directions [7]; nine-bin HOG descriptors summarizing gradient orientation [8]; and shape descriptors such as area, perimeter, circularity, and solidity. Feature-by-sample matrices were transposed to sample-by-feature form before modeling.

2.3. Missing values and normalization

Missing values were handled using a median imputer fitted on the training subset. Two preprocessing modes were compared: raw features and min-max normalization. For normalization, feature minima and maxima were estimated from training data only; validation and test values were clipped to the training range before scaling. This preserved separation between model-development and evaluation data.

2.4. Feature selection

Two feature-selection strategies were evaluated. In t-test-based incremental feature selection (tIFS), features were ranked first by ascending Welch t-test p-value and then, for ties or further ordering, by descending absolute t-statistic [9]. Progressively larger top-k subsets were evaluated. In RF-RFE, a Random Forest was repeatedly fitted on the training subset and, at each iteration, the single feature with the lowest Random-Forest feature importance was removed [10, 11]. Training labels were used for ranking or elimination, whereas validation performance determined the selected feature count. The held-out test set was not used for feature-subset selection.

2.5. Classifiers and model selection

Four classifier families were evaluated in the analyzed run. Logistic regression used solver="liblinear", max_iter=5000, and class_weight="balanced". SVM used an RBF kernel with C=1.0, gamma="scale", and class_weight="balanced" [12]. Random Forest used 100 trees and class_weight="balanced_subsample" [10]. Gradient Boosting used 100 estimators, learning_rate=0.05, and max_depth=3 and did not use class weighting [13]. XGBoost and LightGBM were optional in the code but were unavailable in the analyzed run and are therefore not discussed as evaluated models.

For each feature-selection method, classifier, and preprocessing mode, models were fitted on the training set. Validation AUROC was the primary criterion for selecting feature count; when needed, validation MCC, accuracy, sensitivity, specificity, and then smaller feature count were used as sequential tie-breakers. Among the validation-selected configurations, RF-RFE SVM with min-max normalization and 52 features was designated the representative balanced model because its training and validation metrics were stable and its validation sensitivity and specificity were comparatively balanced. The test set was used only for final independent evaluation and did not determine this designation.

DOI: 10.54254/2753-8818/2026.HZ38953

2.6. Evaluation metrics

Performance was summarized using accuracy, sensitivity, specificity, AUROC, and Matthews correlation coefficient (MCC) [14]. Because high-risk images were the minority class, sensitivity and MCC were interpreted alongside accuracy. AUROC describes ranking performance across possible score thresholds, whereas sensitivity and specificity reflect predictions at the classifier's decision boundary. Confusion matrices were used to inspect false-positive and false-negative patterns.

3. Results

3.1. Dataset composition and initial baselines

The dataset was imbalanced toward low-risk images: high-risk images represented 23.33% of training, 23.81% of validation, and 27.91% of testing. With all 60 features, min-max-normalized SVM achieved validation AUROC 0.6953 with sensitivity 0.6000 and specificity 0.7031, while min-max-normalized logistic regression achieved AUROC 0.6562. Random Forest achieved validation accuracy 0.7500 but sensitivity only 0.0500, illustrating how majority-class performance can inflate accuracy in an imbalanced task.

3.2. Validation-based feature selection and representative model

Validation-based feature selection produced different trade-offs across model families. The highest validation AUROC among selected configurations was 0.7969 for tIFS Random Forest using all 60 features, but its sensitivity was only 0.0500. In contrast, tIFS SVM with seven normalized features achieved sensitivity 0.8000, specificity 0.6250, AUROC 0.7422, and MCC 0.3624; RF-RFE logistic regression with four normalized features achieved sensitivity 0.7000, specificity 0.6875, AUROC 0.7312, and MCC 0.3362. These results demonstrate that high AUROC does not necessarily imply a balanced default decision rule.

The RF-RFE SVM with min-max normalization retained 52 of 60 features and was used as the representative balanced configuration. Table 2 reports the performance of this representative model across the training, validation, and test splits. As shown in Table 2, its MCC values were 0.3311, 0.3086, and 0.2869 across training, validation, and testing, respectively, with similar sensitivity/specificity patterns across splits. This configuration is therefore better interpreted as removing a small number of low-contribution features than as aggressive dimensionality reduction. The retained set included 11 intensity, 6 histogram, 14 GLCM, 9 HOG, and 12 shape features, indicating that useful information was distributed across multiple feature families.

Table 2. Performance of the representative RF-RFE SVM with min-max normalization and 52 features

Split	Accuracy	Sensitivity	Specificity	AUROC	MCC
Training	0.7042	0.6607	0.7174	0.7869	0.3311
Validation	0.7262	0.5500	0.7813	0.7156	0.3086
Testing	0.6977	0.5417	0.7581	0.6969	0.2869

DOI: 10.54254/2753-8818/2026.HZ38953

3.3. Held-out test performance

Table 3 reports selected held-out test results for the final configurations. On the held-out test set, the representative RF-RFE SVM achieved accuracy 0.6977, sensitivity 0.5417, specificity 0.7581, AUROC 0.6969, and MCC 0.2869. Figure 1 presents the corresponding confusion matrix for this representative model. It contained 47 true negatives, 15 false positives, 11 false negatives, and 13 true positives. The RF-RFE Random Forest with min-max normalization achieved the highest test AUROC (0.7382), but sensitivity was only 0.0417, identifying 1 of 24 high-risk images while correctly classifying 61 of 62 low-risk images. Thus, the Random Forest showed stronger continuous-score ranking but a highly unfavorable default decision boundary for the minority class. Figure 2 shows the ROC curves for all validation-selected models on the held-out test set, illustrating moderate separation and variation across configurations.

Table 3. Selected held-out test results. Test performance is reported for independent evaluation, not for model selection. The columns FS, Cls, and Prep give the algorithms for feature selection, classification, and preprocessing. The number of features selected is in the column F. The performance metrics accuracy, sensitivity, specificity, AUROC, and Matthews' correlation coefficient are given in the columns Acc, Sn, Sp, ROC, and MCC

FS	Cls	Prep	F	Acc	Sn	Sp	ROC	MCC
RF-RFE	Random Forest	Min-max	60	0.7209	0.0417	0.9839	0.7382	0.0760
RF-RFE	SVM	Min-max	52	0.6977	0.5417	0.7581	0.6969	0.2869
RF-RFE	Gradient Boosting	Min-max	23	0.6977	0.1667	0.9032	0.6935	0.0978
RF-RFE	Logistic regression	Raw	25	0.6279	0.6250	0.6290	0.6794	0.2294
RF-RFE	Logistic regression	Min-max	4	0.6628	0.5833	0.6935	0.6203	0.2554
tIFS	SVM	Min-max	7	0.6163	0.5000	0.6613	0.6539	0.1488

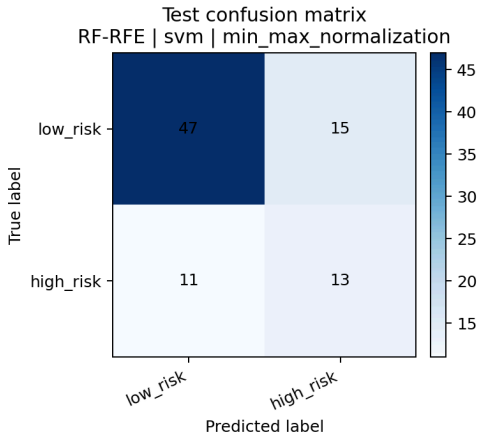

Figure 1. Confusion matrix for the representative RF-RFE SVM with min-max normalization on the held-out test set

DOI: 10.54254/2753-8818/2026.HZ38953

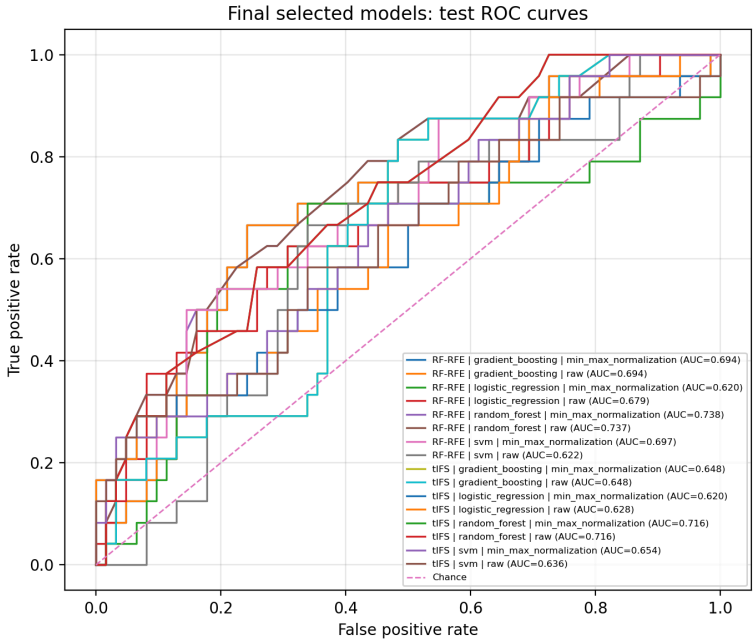

Figure 2. ROC curves for validation-selected models on the held-out test set. Curves show moderate separation and variation across model and feature-selection configurations

3.4. Error patterns

Error analysis identified both false positives and false negatives. False positives may reflect low-risk images with global intensity, texture, or shape patterns resembling high-risk cases, whereas false negatives may reflect subtle or localized high-risk evidence not adequately represented by whole-breast tissue descriptors. Some high-risk images were misclassified across multiple final models, suggesting systematic limitations in the current hand-crafted representation.

4. Discussion

The results show that an interpretable, leakage-aware classical pipeline can extract moderate BI-RADS-associated signal from mammograms, but they also demonstrate why no single metric should determine model interpretation. The representative RF-RFE SVM maintained relatively consistent performance across training, validation, and testing and preserved a more balanced sensitivity-specificity profile. In contrast, Random Forest produced the highest held-out AUROC but extremely low high-risk sensitivity at its default decision boundary. This difference is expected because AUROC evaluates ranking across thresholds, whereas sensitivity and specificity depend on the operating decision rule. The Random Forest pattern, together with its much stronger fit to training data and weak validation/test high-risk detection, is consistent with majority-class bias and an overfitting tendency rather than a practically balanced classifier.

The final feature representation was intentionally transparent but also limited. The representative SVM retained 52 of 60 features spanning all five feature families, so RF-RFE primarily removed a small set of low-contribution variables rather than producing a highly compact model. The broad distribution of retained intensity, histogram, GLCM, HOG, and shape descriptors suggests that the available signal was not concentrated in one feature family. However, because the final matrices were generated from Otsu-derived breast-tissue masks rather than expert lesion ROIs, small lesion-

DOI: 10.54254/2753-8818/2026.HZ38953

specific patterns may have been diluted by global breast structure. This provides a plausible explanation for systematic false negatives and is a central limitation of the current representation.

Normalization particularly benefited scale-sensitive models such as SVM and logistic regression, while tree-based models were less affected. More broadly, the present pipeline should be viewed as an interpretable reference point rather than a competitor to state-of-the-art deep learning. Future learned or multimodal models should be compared using the same patient-level separation and held-out evaluation discipline so that added complexity can be assessed against a transparent baseline.

5. Conclusion

This study developed a reproducible classical machine-learning baseline for BI-RADS-derived mammographic risk classification using 60 interpretable features from Otsu-derived breast-tissue masks. The RF-RFE SVM with min-max normalization and 52 features was selected as the representative balanced configuration based on training/validation stability and balanced validation sensitivity/specificity, and it achieved test sensitivity 0.5417, specificity 0.7581, AUROC 0.6969, and MCC 0.2869. A Random Forest achieved the highest test AUROC (0.7382) but detected only 1 of 24 high-risk images at its default decision boundary, reinforcing the need to interpret AUROC together with sensitivity, specificity, MCC, and confusion matrices. The current outcome is BI-RADS-derived risk classification rather than pathology-confirmed cancer diagnosis. Future work should prioritize lesion-centered features, repeated or external patient-level validation, and comparison with more advanced learned or multimodal models.

6. Limitations

The study has four main limitations. First, the endpoint is derived from BI-RADS categories rather than a complete pathology-confirmed benign/malignant outcome. Second, the dataset is relatively small, imbalanced, and drawn from a single public resource, and the current analysis uses one patient-level partition without external validation. Third, the final hand-crafted features were derived from whole-breast Otsu tissue masks rather than expert lesion-centered ROIs, which may dilute localized abnormalities. Finally, the models were evaluated at their default decision boundaries; threshold selection was not clinically calibrated, so sensitivity and specificity should not be interpreted as a validated clinical operating point.

References

[1] American College of Radiology. (2025). *ACR BI-RADS® atlas: Breast imaging reporting and data system* (2025 ed.). American College of Radiology.

[2] Moreira, I. C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M. J., & Cardoso, J. S. (2012). INbreast: Toward a full-field digital mammographic database. *Academic Radiology, 19*(2), 236–248. https://doi.org/10.1016/j.acra.2011.09.014

[3] Geras, K. J., Mann, R. M., & Moy, L. (2019). Artificial intelligence for mammography and digital breast tomosynthesis: Current concepts and future perspectives. *Radiology, 293*(2), 246–259. https://doi.org/10.1148/radiol.2019182627

[4] Yoon, J. H., & Kim, E.-K. (2021). Deep learning-based artificial intelligence for mammography. *Korean Journal of Radiology, 22*(8), 1225–1239. https://doi.org/10.3348/kjr.2020.1210

[5] Bhalla, D., Rangarajan, K., Chandra, T., Banerjee, S., & Arora, C. (2024). Reproducibility and explainability of deep learning in mammography: A systematic review of literature. *Indian Journal of Radiology and Imaging, 34*(3), 469–487. https://doi.org/10.1055/s-0043-1775737

[6] Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Trans. Syst., Man., Cybern., SMC-9*(1), 62–66. https://doi.org/10.1109/TSMC.1979.4310076

DOI: 10.54254/2753-8818/2026.HZ38953

[7] Haralick, R. M., Shanmugam, K., & Dinstein, I. (1973). Textural features for image classification. *IEEE Trans. Syst., Man., Cybern., SMC-3*(6), 610–621. https://doi.org/10.1109/TSMC.1973.4309314
[8] Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR) (Volume 1)* (pp. 886–893). https://doi.org/10.1109/CVPR.2005.177
[9] Welch, B. L. (1947). The generalization of 'Student's' problem when several different population variances are involved. *Biometrika, 34*(1/2), 28–35. https://doi.org/10.2307/2332510
[10] Breiman, L. (2001). Random forests. *Mach. Learn., 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
[11] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Mach. Learn., 46*(1–3), 389–422. doi: 10.1023/A:1012487302797.
[12] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Mach. Learn., 20*(3), 273–297. doi: 10.1007/BF00994018.
[13] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist., 29*(5), 1189–1232.
[14] Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta (BBA) - Protein Struct., 405*(2), 442–451. doi: 10.1016/0005-2795(75)90109-9.