Zahra: Bulletin Big data, Data Science and Artificial Intelligence
Vol. 3, No. 1, [June 2024], pp. [56–69]
E-ISSN 2830-7984

Research Article Open Access (CC–BY-SA)

Stability-Based Feature Selection and Bayesian Optimization for Interpretable Breast Cancer Classification Using Support Vector Machines

Abdul Karim[1*], Muhammad Syahrizal[2], Dodi Siregar[3], Sadia Samar Ali[4]
[1] *Information Technology Study Program, Faculty of Science and Technology, Universitas Labuhanbatu, Indonesia*
[2] *Computer Graphics Engineering Technology Program, Politeknik Cendana, Indonesia*
[3] *Computer Science Program, Faculty of Engineering and Computer Science, Universitas Harapan, Medan, Indonesia*
[4] *King Abdulaziz University (KAU), Jeddah, Arab Saudi*

[1] *Abdulkarim@ulb.ac.id;* [2] *m.syahrizal156@gmail.com;* [3] *dodi.stth@gmail.com;* [4] *sadiasamarali@gmail.com*
** Corresponding author*

Abstract

Reliable discrimination between malignant and benign breast tumors using diagnostic measurements requires both predictive accuracy and the identification of informative features. This study evaluates whether bootstrap-based stability selection combined with Bayesian hyperparameter optimization can improve Support Vector Machine (SVM) classification while reducing predictor dimensionality. The Wisconsin Breast Cancer Diagnostic dataset comprises 569 observations with 30 quantitative predictors and binary diagnosis labels, with malignant tumors designated as the positive class. A stratified partition allocates 341 observations to training, 114 to model development, and 114 to held-out testing, with feature selection and parameter optimization confined to the development process. Four configurations are compared: a baseline SVM using all predictors, stability-based feature selection alone, Bayesian hyperparameter optimization alone, and the proposed combined stability–Bayesian SVM. The proposed method retains eight predictors, achieving a 73.33% reduction in dimensionality, although its held-out balanced accuracy decreases from 0.9812 for the baseline to 0.9692. Malignant-class sensitivity declines from 0.9762 to 0.9524, while F1 decreases from 0.9762 to 0.9639 and specificity remains unchanged at 0.9861. The proposed model misses two malignant cases compared with one for the baseline, indicating that feature compression introduces a clinically relevant detection trade-off. These findings demonstrate that reducing predictor dimensionality does not necessarily improve classification performance, particularly when the baseline already achieves strong discrimination. The results remain limited by the modest dataset size, a single held-out partition, and the absence of independent clinical validation. Stability-based feature selection and Bayesian optimization should therefore be evaluated through their combined effects on model parsimony, classification sensitivity, and generalizability rather than presumed to provide universal predictive improvement.

Keywords: breast cancer classification; stability selection; Bayesian optimization; support vector machine; balanced accuracy

1. Introduction

Computer-assisted classification of fine-needle aspiration measurements may support assessment of breast lesions, although a benchmark classifier is not a validated clinical diagnostic tool. Stable feature rankings have been investigated as an alternative to choosing predictors from a single model fit [1]. Feature selection combined with Bayesian optimization has also been evaluated for Wisconsin breast cancer classification [2]. Hybrid screening strategies seek compact representations without discarding discriminative information [3]. Image-based feature integration provides a different source of representation learning and cannot be equated directly with cytomorphometric data [4]. Hyperparameter search quality depends on objective definition, validation design, and comparable search budgets [5]. Probability calibration addresses reliability of predicted confidence and should be distinguished from discrimination or an optimized classification threshold [6].

Reducing correlated measurements can make a fitted model simpler, but feature removal can also weaken malignant-case detection. Particle-swarm feature selection has been studied in several breast cancer benchmarks, illustrating the importance of dataset-specific comparisons [7]. Optimization of extracted mammographic features addresses a related representation problem under a different observation modality [8]. Reviews of explainable

diagnostic models identify a need to evaluate the quality and relevance of proposed explanations [9]. Broader evidence on explainable breast cancer systems likewise separates predictive quality from interpretability claims [10]. Empirical multi-omics analyses show that feature-subset stability can depend strongly on the selection method and correlated variables [11]. Comparisons of biomarker-selection algorithms emphasize reproducibility, feature count, and external performance rather than accuracy alone [12].

The present study evaluates stability-based selection and Bayesian SVM optimization using four component-controlled model configurations. Imbalance-aware feature selection has been incorporated into other breast cancer classification frameworks, but those interventions do not establish an advantage for the present hybrid [13]. Explainability reviews highlight that a compact set of variables is not equivalent to a clinically validated biological explanation [14]. Direct Wisconsin-dataset experiments have combined selection and optimization, although their splits and algorithms differ from those evaluated here [15]. Multi-model comparisons illustrate why cross-study accuracy rankings should not substitute for a common held-out comparison [16]. Multimodal diagnostic reviews further caution against generalizing a single tabular benchmark to imaging-based clinical pathways [17]. The objective is to determine whether the combined pipeline improves held-out balanced accuracy while retaining an interpretable subset of measured predictors.

2. Method

2.1. Related Work

Feature stability and model selection are complementary rather than interchangeable objectives. Distance-based selection has been evaluated for breast cancer tabular models, offering an alternative to model-dependent ranking [18]. Resampling and consensus selection have been investigated for biomedical measurements that are sensitive to sample perturbation [19]. Penalized regression experiments on Wisconsin features show that sparse models may have more variable feature inclusion than less sparse alternatives [20]. SVM classification with metaheuristic feature selection has been evaluated on the same diagnostic problem under different validation settings [21]. These findings motivate direct recording of feature-selection frequency alongside held-out sensitivity and specificity. A smaller feature subset is therefore treated as evidence of dimensionality reduction, not automatic evidence of higher reliability.

The broader breast cancer literature provides several methodological comparisons but no transferable guarantee of optimization gains. Explainable KNN studies combine domain-informed features with classical tuning, illustrating that the selected learner and preprocessing jointly affect performance [22]. Modified logistic regression has been benchmarked against SVM and other classifiers while explicitly examining calibration and false-negative errors [23]. Teaching-learning-based optimization has also been used for feature selection in cancer screening benchmarks [24]. Genomic stability selection addresses robustness to contamination and is conceptually related, although its data-generating process differs [25]. Hybrid feature selection on mammography provides evidence about a separate radiomics task rather than tabular FNA data [26]. These studies make common test partitions and complete ablation comparisons more informative than isolated reported maxima.

Predictive feature reduction is frequently pursued to improve transparency and computational efficiency. Multi-omic breast cancer outcome prediction uses optimally selected variables with neural predictors but examines a different endpoint [27]. Breast cancer subtype classification uses distinct biological measurements and does not directly establish benign-versus-malignant screening performance [28]. Systematic syntheses describe variations in breast cancer data modalities, validation, and explanation standards [29]. Stability-aware biomarker commentary warns against treating a high classification score as proof of reliable variable importance [30]. Benchmarking studies of biomarker discovery find that higher model complexity does not uniformly improve accuracy or reproducibility [31]. An informative hybrid evaluation must therefore present sparsity, class-specific errors, and predictive discrimination in the same experimental setting.

Different data modalities also introduce different failure modes and dependencies. Modern mammographic models combine learned multi-scale representations, whereas the present dataset contains structured cytological descriptors [32]. Survival classification with feature reduction targets patient outcomes rather than the diagnostic state of an aspirated breast mass [33]. Feature attribution can be unstable across supervised and unsupervised algorithms even when classification appears accurate [34]. Clinical metaheuristic selection studies highlight the need to distinguish an optimizer's behavior from its eventual interpretability [35]. The Wisconsin benchmark used here comprises a limited number of related nuclear descriptors and no independent institutional cohort. Consequently, classwise false-negative counts and the generalization boundary of a single holdout must remain visible.

The literature supports a controlled test of the individual and combined interventions rather than a presumption of hybrid superiority. A full-feature support vector machine is retained as the principal algorithmic reference. Stability selection alone tests whether feature compression retains predictive quality without Bayesian tuning. Bayesian tuning alone isolates gains attributable to parameter search over the original predictor set. The combined branch tests whether these components interact favorably within the same data partition. Table 1 summarizes comparable contributions and the methodological gap specific to this breast cancer experiment. Its comparison is limited to the reported study designs and does not imply that the methods share datasets, validation budgets, or outcomes.

Table 1. Comparative synthesis of selected related studies

Author / year	Contribution	Research-specific limitation
Yu et al. (2023)	Stable feature ranking for breast cancer classification	Selected-feature stability depends on evaluation design
Akkur et al. (2023)	Feature selection and Bayesian parameter optimization	Different data partitions and learner comparisons
Okyere and Nyanney (2026)	Bootstrap stability and interpretability on WDBC	Penalized regressions rather than a hybrid SVM
Kiani Sarkaleh et al. (2026)	Metaheuristic selection and SVM breast-cancer classification	Different optimizer and experiment settings
Bischl et al. (2023)	Principles of transparent hyperparameter optimization	Methodological guidance rather than a WDBC test
Cattelani et al. (2024)	Feature selection stability and evaluation criteria	Omics biomarker setting; different endpoint

The evidence summarized in Table 1 positions stability-based feature selection alongside SVM tuning and related biomarker methods. The studies address distinct algorithms, data modalities, and validation settings. None establishes that combining stability filtering with Bayesian search guarantees improved Wisconsin classification. The closest methodological comparisons support treating the two interventions as separate factors. Reported improvements from other datasets cannot be transferred to this evaluation without matching the sample allocation and model-selection criterion. The present benchmark therefore assesses the combined and isolated components using one common held-out comparison.

2.2. Dataset

The Wisconsin Diagnostic Breast Cancer dataset comprises 569 records with 30 nuclear-morphometry predictors and a benign-versus-malignant target. The malignant class is treated as positive for sensitivity and false-negative evaluation. The source is the original Kaggle data.csv distributed from the UCI diagnostic collection. Identifier columns and any empty auxiliary fields are excluded from modelling. A stratified split assigns 341 observations to training, 114 to model development, and 114 to the held-out test. Feature selection and all fitted scaling transformations are confined to development-side data. Table 2 summarizes the data provenance and observed audit settings.

Table 2. Planned dataset specification and audit requirements

Dataset attribute	Study-specific protocol
Public source	Kaggle uciml/breast-cancer-wisconsin-data; local data.csv, SHA-256 recorded in execution
Research endpoint	Binary diagnosis; malignant positive, benign negative
Analysis unit	Fine-needle aspiration specimen-level feature record; no independent hospital cohort
Executed partition	Stratified train/development/held-out test = 341/114/114; seed 42
Processing / quality audit	30 numerical predictors; identifier and blank auxiliary fields removed; scaling fitted on development training only
Scope limitation	Single 569-record diagnostic benchmark; no clinical deployment or prospective evaluation

The evidence summarized in Table 2 identifies the dataset and the exact experimental unit used in the evaluation. The response records diagnostic category rather than future disease risk or patient survival. All 569 observations were retained for the reported full-data experiment. The available 30 predictors describe nuclear geometry, texture, and other quantitative cytological properties. The 341/114/114 split prevents routine parameter selection from using

the held-out test observations. The source checksum permits a future researcher to confirm file identity if the original data are provided. The absence of an external cohort limits claims of clinical transferability.

The evidence summarized in Table 3 details the source audit, development operations, and final holdout boundary. The fixed stratified split was generated with seed 42. Standardization was fitted using training observations rather than the independent test. Feature-subset construction and SVM parameter search used the development data. The final test was evaluated only after model variants were frozen. This sequence limits direct data leakage from test labels into selection. The design nevertheless contains only one recorded partition and cannot establish robustness across hospitals or time periods.

Table 3. Experimental data flow and leakage-prevention controls

Stage	Required handling
Stage A: audit	Validate 569 source records, 30 predictors, target mapping and source checksum.
Stage B: split	Stratified 341 train / 114 development / 114 locked test with seed 42.
Stage C: fit	Fit standardization and bootstrap stability selection on training-side data.
Stage D: select	Score full SVM, stability-only, Bayesian-only, and joint candidates using development data.
Stage E: evaluate	Freeze configurations, then compute balanced accuracy, malignant recall, specificity, AUC and confusion counts on 114 tests.

The sequence in Table 3 separates source verification from estimator development and held-out assessment. The input source contains 569 original diagnostic records. Model development uses only the 341 training and 114 development observations. Two Bayesian optimization branches each recorded 50 candidate evaluations. The 114-record final test supports an internally independent comparison of frozen models. The resulting small numbers of malignant errors require exact confusion-count reporting. No evidence of independent equipment, institutions, or prospective patient sampling is provided by this protocol.

2.3. Research Framework

The visual evidence in Figure 1 presents the common data handling shared by the full-feature and stability-based SVM branches. The default classifier uses all 30 standardized diagnostic predictors. The proposed branch selects a stable subset before Bayesian parameter search. A stability-only branch tests feature reduction without Bayesian tuning. A Bayesian-only branch tests parameter search while retaining all predictors. All four decisions use the same source observations and locked final test. This arrangement distinguishes the contribution of feature filtering from that of hyperparameter tuning.

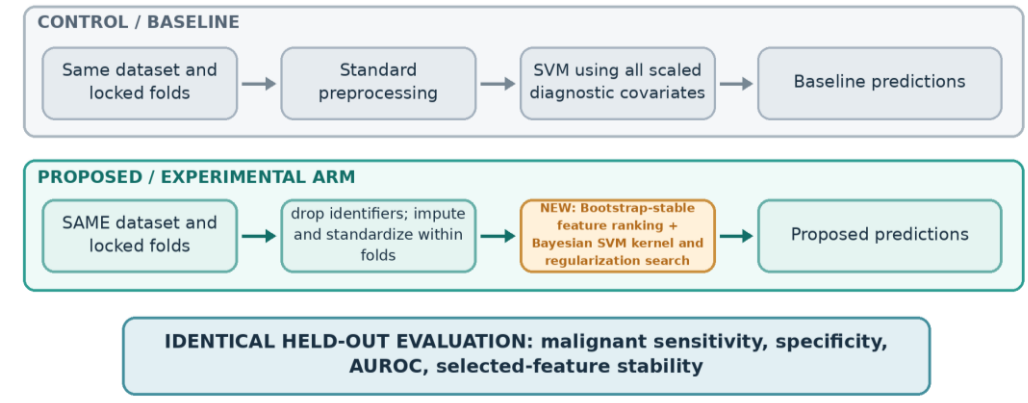

Figure 1. Controlled baseline-versus-proposed research framework

The visual evidence in Figure 1 isolates the algorithmic changes relative to the full-feature baseline. The baseline operates on 30 numerical predictors after standardization. Stability filtering retains eight variables in its reduced-feature branch. The Bayesian search evaluates SVM configurations using development data, not final-test outcomes. Using the same final set makes false-negative and false-positive counts directly comparable. The diagram describes the experimental contrasts, while the final comparison determines whether the combined intervention actually improves performance.

The evidence summarized in Table 4 defines four separate experimental treatments built from the same SVM family. The full-feature baseline provides a standard reference. Stability-only estimation tests feature selection

without Bayesian parameter search. Bayesian-only tuning tests the contribution of search without dimensionality reduction. The combined procedure evaluates both adjustments on the eight-feature subset. All four configurations retain the same malignant-positive label and final test samples. The table precedes the formal definitions of stability selection and the classification scores.

Table 4. Baseline, proposed intervention and falsification design

Component	Comparative specification
Baseline	Standardized full-feature SVM (30 predictors).
Proposed innovations	Bootstrap stability-selected eight-feature subset + Bayesian SVM optimization.
Controlled factors	Common 341/114/114 split, seed 42, positive class, evaluation code and held-out cases.
Executed ablations	All-feature baseline; stability-only; Bayesian-only; stable+Bayesian hybrid.
Primary comparisons	Test balanced accuracy; malignant recall; specificity; F1; MCC; ROC-AUC; selected-feature count.

The evidence summarized in Table 4 makes the component-level comparisons explicit. Feature selection and SVM parameter optimization are tested separately and together. Both Bayesian branches use 50 recorded development objective evaluations. The baseline and stability-only branches do not require matching search trials because no Bayesian search is applied. The primary endpoint is balanced accuracy rather than overall accuracy alone. A superior hybrid would need to outperform the strong full-feature reference under this common test. The recorded experiment does not satisfy that directional criterion.

2.4. Proposed Method

The visual evidence in Figure 2 depicts bootstrap-stable feature selection followed by Bayesian SVM parameter optimization. A selection frequency is computed for each candidate feature across training-side resamples. Features meeting the chosen stability criterion define the reduced design matrix. SVM kernel and regularization settings are then selected using development data. The resulting estimator produces malignant-versus-benign predictions on the held-out cases. The same baseline and component-only controls remain available for the final comparison. Figure 2 illustrates this intended design without implying that lower dimensionality guarantees greater diagnostic accuracy.

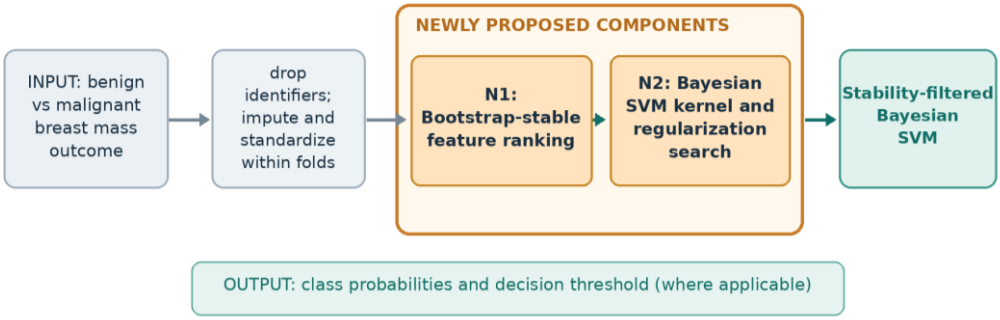

Figure 2. Proposed method architecture with explicitly highlighted novel components

The proposed approach combines feature resampling and model tuning as illustrated in Figure 2. Selection frequency quantifies how consistently a variable enters the bootstrap-derived subset. The retained set contains eight of the 30 original measurements in the executed comparison. Model fitting and SVM parameter search are separated from the locked test evaluation. Bayesian optimization supplies sequential candidate selection rather than an exhaustive parameter grid. The combined method is assessed against feature selection alone and Bayesian tuning alone. Its output remains a binary predictive classification rather than an explanation of biological causation.

The evidence summarized in Table 5 provides the mathematical and operational components of the stability-filtered Bayesian SVM. The bootstrap inclusion frequency describes stability of a candidate feature. Its threshold produces a selected feature subset from the training observations. Standardization is learned without using held-out samples. Candidate SVM settings are compared through development balanced accuracy. The experimental contrast includes a full-feature baseline and the two single-component ablations. The table is followed by editable mathematical definitions used to interpret the reported outcomes.

Table 5. Proposed methodological components and verification rules

Design element	Operational definition
Standard input branch	Drop ID and unnamed column; standardize numerical features within fitted training pipeline.
Feature-selection component	Bootstrap-resampled feature inclusion frequencies, with eight retained predictors.
SVM optimization component	Bayesian search on development side; 50 evaluations per Bayesian-only and joint branch.
Output model	Binary SVM classifier; malignant-positive labels.
Decision / objective	Maximize development balanced accuracy; stable subset determined from training-side frequencies.
Executed ablation	Full-feature baseline, stability-only, Bayesian-only, stable+Bayesian hybrid.
Evaluation endpoints	Balanced accuracy, malignant recall, specificity, F1, MCC, ROC-AUC, PR-AUC and subset size.

The evidence summarized in Table 5 separates feature inclusion from SVM hyperparameter search and from the final classification rule. The frequency of inclusion across bootstrap resamples is a stability statistic, not a model accuracy score. The selected subset is determined before fitting the proposed final classifier. Development balanced accuracy guides the reported Bayesian search rather than the final test. Candidate SVM kernels and regularization values are implementation dependent. The test comparison uses identical malignant-positive definitions and the same number of records. The subsequent equations formalize these concepts without introducing unverified optimizer penalties.

The recorded full-data experiment uses 569 observations and seed 42. A stratified allocation provides 341 training, 114 development, and 114 held-out test records. Full-feature and reduced-feature SVM variants share the same test cases and label mapping. Each Bayesian branch has 50 recorded objective evaluations, for a total of 100 search calls. The primary endpoint is balanced accuracy, complemented by class-sensitive and ranking metrics. Training-only stability frequencies and post-hoc permutation importance have different inferential meanings. The experiment does not establish external validation, repeated-split variability, or a causal meaning for selected biomarkers.

Equations (1)-(7) formalize the main feature-selection and classification quantities. Equation (1) gives the bootstrap inclusion frequency for a predictor. Equation (2) defines the retained stable subset under an inclusion threshold. Equations (3) and (4) express development-based SVM selection and the binary decision rule. Equation (5) defines balanced accuracy using sensitivity and specificity. Equations (6) and (7) provide the observed feature reduction and held-out primary-metric difference. The mathematical expressions are evaluation definitions and do not assert an unverified optimizer-internal loss penalty.

$$S_j = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\{j \in \mathcal{S}^{(b)}\} \tag{1}$$

$$\mathcal{S}_\pi = \{j : S_j \geq \pi_{\text{thr}}\} \tag{2}$$

$$h^\star = \arg\max_{h \in \mathcal{H}} BA_{\text{dev}}(\text{SVM}_h) \tag{3}$$

$$\hat{y}_h(x) = \text{sign}\big(f_h(x)\big) \tag{4}$$

$$BA = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{5}$$

$$FR(\%) = 100\,\frac{30-8}{30} = 73.33\% \tag{6}$$

$$\Delta BA = BA_{\text{joint}} - BA_{\text{baseline}} = 0.969246 - 0.981151 = -0.011905 \tag{7}$$

The definitions distinguish reproducible feature inclusion from predictive discrimination on unseen records. The frequency statistic is measured within the available training-side resampling procedure. The SVM prediction rule operates on the standardized feature representation. The confusion-based balanced-accuracy equation gives equal weight to malignant sensitivity and benign specificity. Feature reduction uses the original 30 columns as the

denominator. The observed difference is negative because the hybrid's test balanced accuracy is below the full-feature baseline. All quantities remain descriptive of the executed internal comparison.

3. Results and Discussion

3.1. Data Distribution and Development Selection

The executed dataset contains 569 diagnostic observations, each represented by 30 numerical breast-mass descriptors. The positive class is malignant and the negative class is benign. A stratified random seed of 42 defines 341 training records, 114 development records, and 114 locked-test records. Figure 3 displays missingness and the class distribution from the source audit. The displayed feature columns have no recorded missing observations. The modest sample size makes the interpretation of one or two additional errors consequential. Class-specific results must therefore accompany aggregate metrics. The source overview in Figure 3 summarizes measured missingness and the binary diagnosis balance. The independent variable space contains 30 recorded numeric descriptors. The audit covers the entire 569-record dataset. The class counts explain why the malignant-positive rate must be kept explicit. The graph is descriptive rather than a performance evaluation. Subsequent analysis uses the same records for all four models.

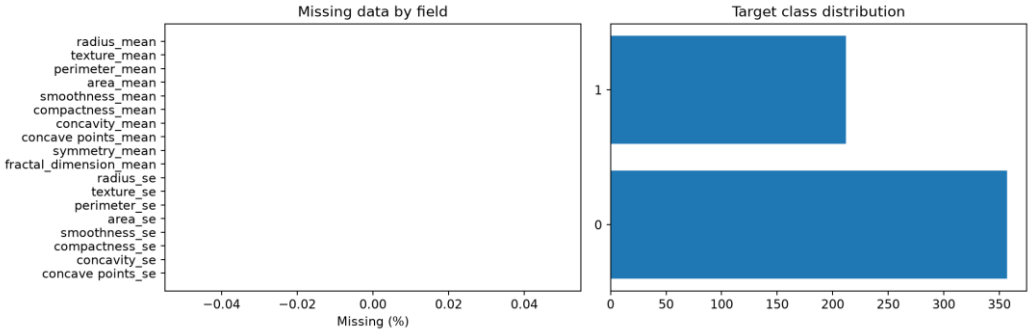

Figure 3. Original-data diagnosis distribution and feature missingness

The visual evidence in Figure 3 illustrates the binary distribution and the absence of measured missing values in the inspected predictors. The smaller malignant group carries the positive classification label. This imbalance is moderate rather than extreme and does not justify interpreting high accuracy as sufficient by itself. A missed malignancy has different clinical meaning from a false alarm, although no patient-level cost ratio was measured here. The figure supports the use of malignant sensitivity together with specificity and balanced accuracy. The chart does not establish independence across institutions or future patients.

The evidence summarized in Table 6 reports the development scores that determined each candidate's configuration before the locked test. The baseline and stability-only branches were fitted without Bayesian search. Two Bayesian search branches each used 50 recorded objective evaluations. Development balanced accuracy is the reported selection criterion. The feature counts differ between full and stability-filtered models. The held-out comparison subsequently tests whether the development ranking generalizes. These observations distinguish selection performance from test evidence.

Table 6. Development-stage model selection and objective budgets

Model	Development BA	Predictors	BO trials
Full-feature SVM	0.974288	30	--
Stability-only SVM	0.962660	8	--
Bayesian-only SVM	0.992958	30	50
Hybrid stability + Bayesian SVM	0.981330	8	50

The evidence summarized in Table 6 shows that Bayesian-only tuning achieved the highest observed development balanced accuracy. The hybrid ranked below Bayesian-only tuning despite retaining a compact eight-feature subset. Stability selection alone was less accurate on development data than the full-feature baseline. These development scores were used to identify model configurations before held-out testing. The 50 evaluations in each optimized branch are recorded search calls, not repeated independent experiments. The test comparisons are therefore required to judge predictive benefit without development-selection optimism.

3.2. Held-Out Classification Performance

The evidence summarized in Table 7 presents the final held-out baseline and ablation results under identical 114-observation evaluation conditions. Malignant diagnosis is the positive class for the sensitivity and precision measures. Balanced accuracy is the primary endpoint and equally weights malignant sensitivity and benign specificity. The table also includes standard accuracy, F1, MCC, feature count, and confusion counts. All scores were obtained with configurations fixed from development data. The table permits a direct assessment of whether the proposed hybrid adds predictive value. The evidence does not support a superiority claim for the combined method.

Table 7. Locked-test SVM baseline versus stability and Bayesian optimization

Indicator	Baseline SVM	Stable only	Bayesian only	Hybrid
Predictors	30	8	30	8
Accuracy	0.982456	0.973684	0.973684	0.973684
Balanced accuracy	0.981151	0.969246	0.969246	0.969246
Malignant precision	0.976190	0.975610	0.975610	0.975610
Malignant recall	0.976190	0.952381	0.952381	0.952381
Benign specificity	0.986111	0.986111	0.986111	0.986111
Malignant F1	0.976190	0.963855	0.963855	0.963855
MCC	0.962302	0.943340	0.943340	0.943340
TP / FN	41 / 1	40 / 2	40 / 2	40 / 2
TN / FP	71 / 1	71 / 1	71 / 1	71 / 1

The four-way comparison in Table 7 identifies the full-feature baseline as the best model on the prespecified balanced-accuracy endpoint. Its score of 0.981151 exceeds the hybrid's 0.969246 by 0.011905. Both models correctly classify 71 benign cases and misclassify one benign case as malignant. The hybrid correctly detects 40 malignant records, compared with 41 for the baseline. The extra missed malignancy lowers recall and F1 while specificity remains unchanged. Stability-only and Bayesian-only variants match the hybrid's final confusion counts. The hybrid therefore achieves dimensionality reduction without any held-out predictive improvement.

The visual evidence in Figure 4 presents the complete set of model-level held-out scores. The figure derives from the same frozen test predictions used in Table 7. It juxtaposes aggregate classification metrics instead of emphasizing one favorable number. Feature count is interpreted separately from predictive quality. All four models share the same malignant-positive label. The model-level plot is introduced here before its visual evidence. The figure does not represent a new evaluation split.

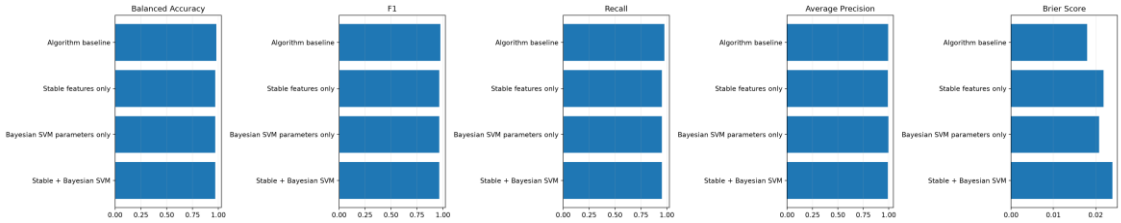

Figure 4. Held-out classification metrics across four SVM branches

The visual evidence in Figure 4 shows that the default full-feature classifier retains the highest balanced accuracy and F1. The three adjusted configurations obtain matching confusion-based metrics on the observed test records. Feature reduction therefore does not translate to better aggregate classification. Small visual differences among threshold-independent measures should not be reinterpreted as superiority on the primary endpoint. The test contains only 42 malignant observations, so one additional missed case changes recall materially. These results reinforce the need to report class-specific counts together with the numerical scores.

The visual evidence in Figure 5 displays the four confusion matrices from the locked test. All matrices use 42 malignant and 72 benign labels. The positive class is malignant rather than benign. Each matrix records true and false classification counts at the frozen operating rule. This visualization enables direct inspection of the single additional malignant error in the adjusted models. It also confirms the common test denominator. The figure precedes discussion of the sensitivity trade-off.

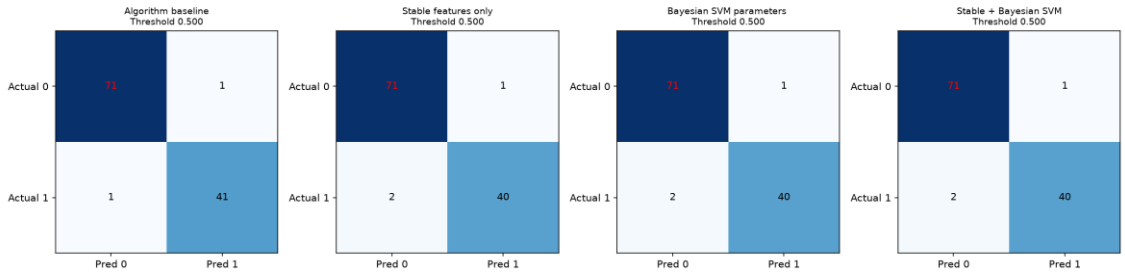

Figure 5. Locked-test confusion matrices for baseline and optimized classifiers

The matrices in Figure 5 show that the baseline produces 41 true positives, one false negative, 71 true negatives, and one false positive. Each of the other three models produces 40 true positives, two false negatives, 71 true negatives, and one false positive. This result explains the observed balanced-accuracy decrease without relying on a rounded percentage. Benign specificity is identical across all four branches. The extra missed malignant record is consequential because malignant detection is a central aim of the study. The visual confirms that the hybrid has not improved the principal error trade-off on this test.

The visual evidence in Figure 6 presents receiver operating characteristic curves from the final predictions. ROC-AUC evaluates the ranking of predicted scores across possible thresholds. It is distinct from balanced accuracy at the frozen decision rule. The curves are calculated from the same held-out labels for each fitted model. Minor ranking differences may occur even when two configurations have identical confusion matrices. The figure is included as a secondary discrimination diagnostic. Threshold-independent performance cannot overturn the prespecified primary result.

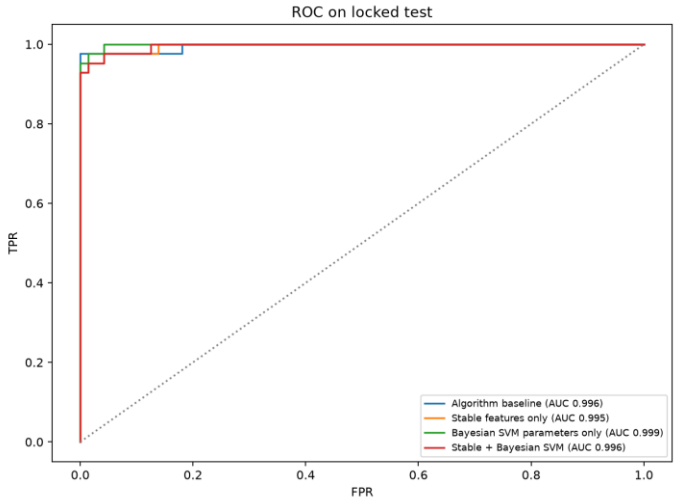

Figure 6. ROC curves for SVM models on the locked test

The visual evidence in Figure 6 shows strong discrimination across the compared SVM classifiers on this internal holdout. The curves lie predominantly near the upper-left region of ROC space. The similar shape of the curves is consistent with high performance on a small established diagnostic benchmark. ROC-AUC summarizes ranking quality rather than the exact number of malignancies missed at the selected threshold. Consequently, the result does not negate the one-case recall deficit of the hybrid. No external cohort was assessed to determine whether the ranking patterns would transfer to new clinical settings.

The visual evidence in Figure 7 provides precision-recall curves for the malignant-positive class. This presentation is sensitive to the relationship between detected malignancies and false-positive alarms. It complements ROC analysis by emphasizing positive-class retrieval. Each curve uses the same 114 held-out observations. Precision-recall behavior is presented separately from feature-selection stability. The diagnostic objective includes reducing missed positive cases. The displayed curves therefore inform, but do not replace, the confusion-based analysis.

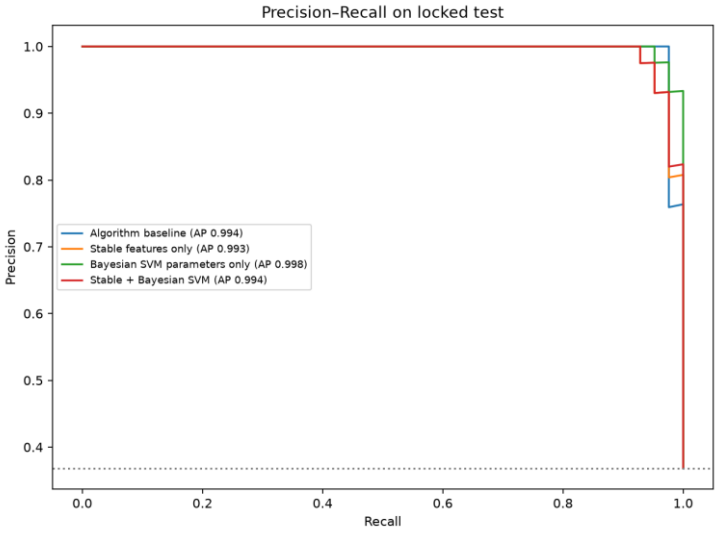

Figure 7. Malignant-positive precision-recall curves

The curves in Figure 7 indicate high precision over much of the observed recall range. Precision decreases at more permissive operating points because additional benign cases are classified as malignant. The hybrid does not show an established increase in detected malignant cases at its frozen operating rule. Its measured sensitivity is lower than that of the full-feature baseline. Curves based on 42 positive test cases are sensitive to individual observations. The visual supports a qualified discrimination assessment rather than a clinical screening claim.

3.3. Feature Stability, Optimization and Reliability

The visual evidence in Figure 8 displays the frequency of feature inclusion across training-side bootstrap resamples. The frequency statistic is conceptually different from a feature's causal or clinical importance. A dashed line indicates the selection criterion used in the visualized stability analysis. The reduced pipeline retains eight candidate predictors. High selection frequency suggests consistency under the recorded resampling procedure. The source feature names are preserved directly from the experimental output. The figure supports examination of variable reproducibility without implying independent validation.

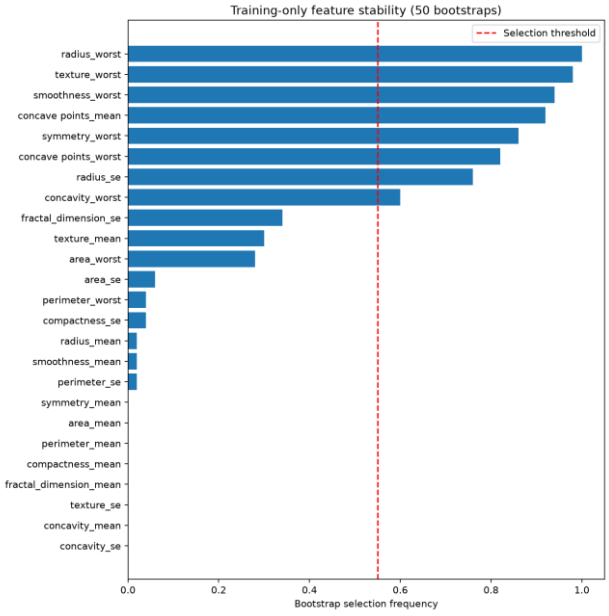

Figure 8. Bootstrap feature-selection frequencies and selected subset

The visual evidence in Figure 8 ranks the original predictors by how often they were selected during training resampling. Measurements associated with tumor size, contour, and surface characteristics appear among the more frequently retained candidates. The final hybrid subset contains eight of 30 available predictors. This corresponds to

a reduction of 73.33% in input dimensionality. Stability within resamples does not establish biomarker causality or selection stability in a different population. The selected features should therefore be interpreted as an internally stable predictive subset rather than validated clinical factors.

The visual evidence in Figure 9 presents the recorded development optimization search trajectory. The plot includes the candidate objective evaluations from both Bayesian branches. The horizontal axis counts completed evaluations rather than independent test replicates. Candidate scores were evaluated using development data, not the locked test. The search trace is useful for distinguishing explored settings from selected finalists. Its values should not be interpreted as confidence intervals around generalization performance. The graph documents optimization behavior without altering the final model ranking.

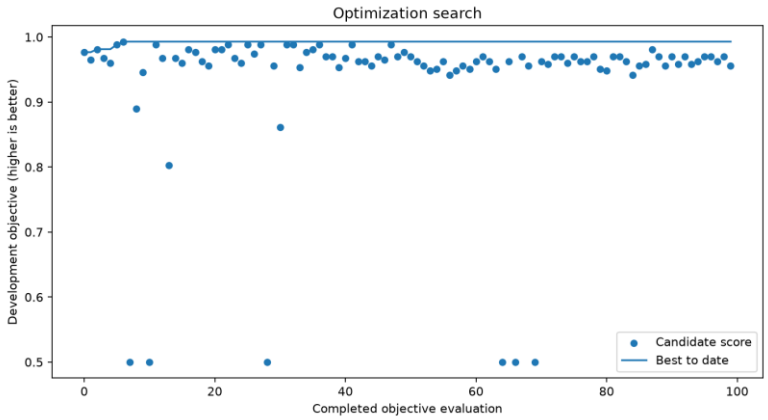

Figure 9. Recorded Bayesian optimization objective evaluations

The visual evidence in Figure 9 shows variability in development objective values across candidate evaluations. Both Bayesian branches account for 50 evaluations each, totaling 100 recorded calls. Some candidates score substantially below the better configurations, indicating that not all parameter settings were competitive. A high development score did not guarantee the strongest held-out balanced accuracy. The baseline SVM remained the best observed final-test configuration. The trace illustrates the distinction between search efficiency and predictive generalization, not a causal explanation for test errors.

The visual evidence in Figure 10 displays observed probabilities against outcome frequency in the final sample. Reliability curves evaluate whether empirical event proportions are aligned with predicted risk bins. Such diagnostics complement discrimination and hard-label classification metrics. The plot is generated from the original locked-test probabilities. Only 114 test records are available for estimating calibration bins. Sparse bins can therefore produce visually unstable reliability estimates. The figure is interpreted as an exploratory diagnostic rather than evidence of deployment-grade calibration.

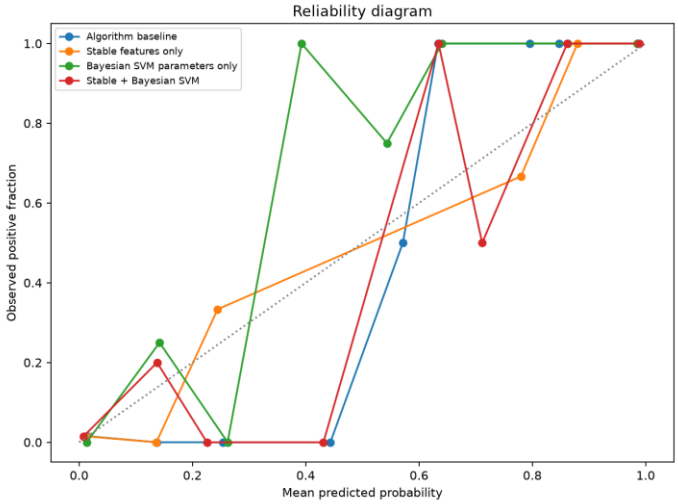

Figure 10. Reliability diagrams for SVM predictions on the final test

The visual evidence in Figure 10 illustrates departures from the diagonal perfect-calibration reference at several probability intervals. Apparent differences across methods reflect both predicted probabilities and small bin sample sizes. The plot does not demonstrate that independent calibration procedures were fitted to each branch. High classification accuracy can coexist with poorly reliable probability levels. These observations counsel against applying the uncalibrated scores directly to patient-level decisions. Independent calibration assessment on a larger external cohort would be necessary before clinical risk interpretation.

3.4. Interpretation and Limitations

The proposed combination primarily reduces model input dimensionality rather than increasing predictive performance. Removing 22 of 30 descriptors may simplify feature acquisition or presentation, but those benefits were not measured in clinical practice. The baseline's higher balanced accuracy arises from one fewer missed malignant case at unchanged benign specificity. Bayesian-only tuning and stability-only selection produce the same held-out confusion counts as the hybrid. The sample contains only 114 test records, and a single partition does not quantify the stability of model rankings under repeated sampling. Feature frequencies describe resampling consistency, not causality or external transportability. No real diagnostic impact, measured clinical costs, or multi-institution validation can be inferred from this benchmark.

4. Conclusion

Four SVM variants were evaluated with a 341/114/114 split of 569 diagnostic records. The baseline attained held-out balanced accuracy of 0.981151. The hybrid achieved 0.969246 with eight rather than 30 predictors. Feature reduction was 73.33%, but malignant recall declined from 0.976190 to 0.952381. The hybrid missed one more malignant case at unchanged benign specificity. Neither single-component intervention outperformed the baseline. The combined method therefore reduced dimensionality without a predictive advantage. The findings distinguish fewer predictors from better malignancy detection. Bootstrap frequencies describe internal subset consistency, not biological causality. A single partition leaves the ranking of methods uncertain. The absence of external clinical cohorts also limits transportability. Future investigations should test repeated splits and paired uncertainty intervals. Independent calibration and external validation are required before clinical application.

References

[1] S. Yu, M. Jin, T. Wen, and others, “Accurate breast cancer diagnosis using a stable feature ranking algorithm,” *BMC Med. Inform. Decis. Mak.*, 2023, doi: 10.1186/s12911-023-02142-2.

[2] E. Akkur, F. Turk, and O. Erogul, “Breast Cancer Diagnosis Using Feature Selection Approaches and Bayesian Optimization,” *Computer Systems Science and Engineering*, 2023, doi: 10.32604/csse.2023.033003.

[3] W. M. Shaban, “Insight into breast cancer detection: new hybrid feature selection method,” *Neural Comput. Appl.*, 2023, doi: 10.1007/s00521-022-08062-y.

[4] A. M. Hassan, A. Yahya, and A. Aboshosha, “A framework for classifying breast cancer based on deep features integration and selection,” *Neural Comput. Appl.*, 2023, doi: 10.1007/s00521-023-08341-2.

[5] B. Bischl, M. Binder, M. Lang, and others, “Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges,” *WIREs Data Mining and Knowledge Discovery*, 2023, doi: 10.1002/widm.1484.

[6] T. de Menezes e Silva Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, and P. Flach, “Classifier calibration: A survey on how to assess and improve predicted class probabilities,” *Mach. Learn.*, 2023, doi: 10.1007/s10994-023-06336-7.

[7] R. Kazerani, “Improving Breast Cancer Diagnosis Accuracy by Particle Swarm Optimization Feature Selection,” *International Journal of Computational Intelligence Systems*, 2024, doi: 10.1007/s44196-024-00428-5.

[8] S. Chakravarthy, B. Nagarajan, V. V Kumar, and others, “Breast Tumor Classification with Enhanced Transfer Learning Features and Selection Using Chaotic Map-Based Optimization,” *International Journal of Computational Intelligence Systems*, 2024, doi: 10.1007/s44196-024-00409-8.

[9] D. K. Gurmessa and W. Jimma, “Explainable machine learning for breast cancer diagnosis from mammography and ultrasound images: a systematic review,” *BMJ Health Care Inform.*, 2024, doi: 10.1136/bmjhci-2023-100954.

[10] A. Ghasemi, S. Hashtarkhani, D. L. Schwartz, and A. Shaban-Nejad, “Explainable artificial intelligence in breast cancer detection and risk prediction: A systematic scoping review,” *Cancer Innovation*, 2024, doi: 10.1002/cai2.136.

[11] T. Lukaszuk, J. Krawczuk, K. Zyla, and J. Kesik, "Stability of Feature Selection in Multi-Omics Data Analysis," *Applied Sciences*, 2024, doi: 10.3390/app142311103.

[12] L. Cattelani, A. Ghosh, T. J. Rintala, and V. Fortino, "A Comprehensive Evaluation Framework for Benchmarking Multi-Objective Feature Selection in Omics-Based Biomarker Discovery," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 2024, doi: 10.1109/TCBB.2024.3480150.

[13] I. Chhillar and A. Singh, "An improved soft voting-based machine learning technique to detect breast cancer utilizing effective feature selection and SMOTE-ENN class balancing," *Discover Artificial Intelligence*, 2025, doi: 10.1007/s44163-025-00224-w.

[14] Z. A. Ansari, M. M. Tripathi, and R. Ahmed, "The role of explainable AI in enhancing breast cancer diagnosis using machine learning and deep learning models," *Discover Artificial Intelligence*, 2025, doi: 10.1007/s44163-025-00307-8.

[15] J. Zhu *et al.*, "An integrated approach of feature selection and machine learning for early detection of breast cancer," *Sci. Rep.*, 2025, doi: 10.1038/s41598-025-97685-x.

[16] X. Zhang, W. Shao, M. Qiu, C. Xiao, and L. Ma, "Advanced deep learning and transfer learning approaches for breast cancer classification using advanced multi-line classifiers and datasets with model optimization and interpretability," *PeerJ Comput. Sci.*, 2025, doi: 10.7717/peerj-cs.2951.

[17] D. Alzamil, B. Alkhamees, and M. M. Hassan, "A Systematic Review of Multimodal Fusion and Explainable AI Applications in Breast Cancer Diagnosis," *Computer Modeling in Engineering and Sciences*, 2025, doi: 10.32604/cmes.2025.070867.

[18] S. S. Baawi, M. N. Kadhim, and D. Al-Shammary, "Efficient feature selection based on Gower distance for breast cancer diagnosis," *Journal of Electronic Science and Technology*, 2025, doi: 10.1016/j.jnlest.2025.100315.

[19] Z. Liu, J. Du, J. Luo, Y. Yuan, Q. Huang, and J. He, "A stable feature selection method based on majority voting and SHAP for high-dimensional metabolomics data," *Comput. Methods Programs Biomed.*, 2026, doi: 10.1016/j.cmpb.2025.109170.

[20] F. Okyere and M. Nyanney, "Stability and interpretability of penalized logistic regression models for breast cancer risk prediction," *PLoS One*, 2026, doi: 10.1371/journal.pone.0353489.

[21] M. Kiani Sarkaleh, H. Azgomi, and A. Kiani-Sarkaleh, "Breast Cancer Classification Using Feature Selection via Improved Simulated Annealing and SVM Classifier," *Diagnostics*, 2026, doi: 10.3390/diagnostics16040637.

[22] O. AlOmair, O. M. Zakaria, M. Alabdullatif, and Z. Khurshid, "Optimized KNN with domain-informed features and LIME explainability for improved breast cancer classification," *BMC Med. Inform. Decis. Mak.*, 2026, doi: 10.1186/s12911-026-03427-y.

[23] A. Sudhagar, S. Athithan, V. Sivakumar, D. Datta, and S. Kaliraj, "A novel approach to breast cancer detection using modified logistic regression with optimized feature selection," *Sci. Rep.*, 2026, doi: 10.1038/s41598-026-64960-4.

[24] A. K. Shukla, S. Dwivedi, S. K. Singh, and others, "High-performance next-generation screening technology system using teaching-learning-based optimization for cancer detection," *Discover Computing*, 2026, doi: 10.1007/s10791-026-10450-0.

[25] G. Yang, X. Lu, and C. Wu, "Robust prioritization of genomic features with stability selection," *Bioinformatics*, 2026, doi: 10.1093/bioinformatics/btag398.

[26] B. S. Alshreef, "Hybrid metaheuristic feature selection for breast cancer detection in digital mammography: a radiomics and deep learning pilot feasibility study," *BMC Med. Imaging*, 2026, doi: 10.1186/s12880-026-02507-9.

[27] P. Tamrakar and G. R. Pathak, "Random optimization-based feature selection for breast cancer outcome prediction using fully connected neural networks on multi-omic data," *Discover Artificial Intelligence*, 2026, doi: 10.1007/s44163-026-01205-3.

[28] A. B. Miranda Valentin, G. M. Bressan, and E. A. da Silva Lizzi, "Unsupervised and supervised approaches for breast cancer subtype classification: hierarchical clustering and machine learning with hyperparameter optimization," *Research on Biomedical Engineering*, 2026, doi: 10.1007/s42600-026-00484-0.

[29] J. Owotogbe, E. Oyekanmi, S. E. Adepoju, and A. E. Akinsunmade, "Machine learning and deep learning for breast cancer: A decade systematic review of detection, classification, prognosis, and explainability," *Inform. Med. Unlocked*, 2026, doi: 10.1016/j.imu.2026.101756.

[30] Y. Takefuji, "Correspondence: Accuracy Is Not Enough: Stability-Aware Feature Selection for Reproducible Biomarker Discovery," *Allergy*, 2025, doi: 10.1111/all.70075.

[31] C. M. Njume, I. Petracci, S. Bellini, and others, “When complexity does not pay: benchmarking deep learning and ensemble methods for biomarker discovery,” *Brief. Bioinform.*, 2026, doi: 10.1093/bib/bbag211.

[32] K. Anusha and K. Reddy Madhavi, “AMFFR-Net: Adaptive Multi-scale Feature Fusion and Graph Attention Networks for Breast Cancer Classification in Mammography Images,” *International Journal of Computational Intelligence Systems*, 2026, doi: 10.1007/s44196-026-01168-4.

[33] S. S. Tharadheesh, S. Aswin, P. Manikandan, and C. Ponnuraja, “Enhanced breast cancer survival status classification through MI-PCA based feature selection and stacking ensemble learning,” *Array*, 2026, doi: 10.1016/j.array.2026.101043.

[34] Y. Takefuji, “Stability of feature attribution: Contrasting supervised and unsupervised selection for radiopathomics and TCGA outcomes,” *Cancer Lett.*, 2026, doi: 10.1016/j.canlet.2025.218133.

[35] S. Adamu, H. Alhussian, S. J. Abdulkadir, and others, “Interpretable hybrid metaheuristic optimization with iteration-level behavior analysis for clinical feature selection in heart disease prediction,” *Journal of King Saud University Computer and Information Sciences*, 2026, doi: 10.1007/s44443-026-00868-3.