

Ensemble Feature selection with Adaptive Weighted voting for Robust Breast Cancer Diagnosis

Muhammad Umair^{1,3}, Hafiz Muhammad Tayyab Khushi^{1,2*}, Tehreem Masood^{1,3}, Iftikhar Naseer^{1,4}, Shamim Akhter⁵, and H. M. Shahzad^{1,3}

¹Faculty of Computer Science and Information Technology, The Superior University, Lahore, 54000, Pakistan.

²Department of Information Technology, The Superior University, Lahore, 54000, Pakistan.

³Department of Software Engineering, The Superior University, Lahore, 54000, Pakistan.

⁴Superior College, Mian Channu, 58000, Pakistan.

⁵School of Information Management, Minhaj University 54600, Pakistan.

*Corresponding Author: Hafiz Muhammad Tayyab Khushi. Email: muhammad.tayyab@superior.edu.pk

Received: May 20, 2026 Accepted: September 17, 2026

Abstract: Breast cancer is a significant disease prevalent in many women around the globe and one of the most common causes of cancer death worldwide, this underscores the demand for trustworthy and precise computer-aided diagnostic tools. Typically, high-dimensional medical datasets possess redundant and irrelevant features which may adversely affect the classification performance and explainability of models. Moreover, conventional feature selection methods frequently suffer from instability, selecting different feature subsets across training samples, which limits reproducibility in clinical decision-making. To address these challenges, this study proposes an adaptive ensemble feature selection framework that integrates four complementary feature selection methods—namely Mutual Information (MI), ANOVA F-test, LASSO, and Random Forest importance—using three ensemble Voting strategies: hard majority Voting, performance-weighted voting, and rank aggregation. The adaptive weights are estimated using balanced accuracy from 5-fold cross-validation in the training data to ensure data-driven performance and to avoid information leakage during the feature selection process. The proposed adaptive ensemble framework is tested on three publicly available breast cancer datasets based on different clinical domains, namely the Coimbra dataset (blood-based biomarkers), the SEER dataset (epidemiological records), and the Mammographic Mass dataset (BI-RADS characteristics). As a result of the experiment, the framework achieves the balanced accuracies of 88.46%, 78.00%, and 84.77% on the Coimbra, SEER, and Mammographic Mass datasets, correspondingly, and perfect feature stability (Jaccard Index = 1.0) for all datasets in repeated cross-validation. Moreover, the selected features are clinically meaningful, including such diagnostic parameters as glucose, BMI, resistin, tumor characteristics, and BI-RADS characteristics. This means that the adaptive ensemble framework allows obtaining an accurate, stable, and interpretable feature selection algorithm that can be used for decision support in breast cancer diagnosis.

Keywords: Ensemble Feature Selection; Breast Cancer Diagnosis; Adaptive Weighted Voting; Feature Stability; Balanced Accuracy; Machine Learning; Clinical Decision Support

1. Introduction

Breast cancer is now the most common type of cancer among women and is the second most common cause of deaths due to cancer around the world [1]. According to World Health Organization there are 2.3 million new cases of breast cancer in 2012 and 670,000 deaths were caused due to breast cancer in 2012. Early diagnosis provides promising evidence for improved survival for breast cancer. However, the current procedures used for diagnosis are plagued with inaccuracies, cost and accessibility issues [2], [3].

Over recent times, machine learning has become increasingly significant within health care data analysis. Using clinical information, patient notes, and scans, such methods train accurate forecasting tools alongside diagnosing systems. However, because medical information often brings large volumes of variables into play - resulting in issues like growing complexity, poor generalized performance, or hard-to-understand outputs - careful handling of inputs may be necessary. That is where selective choice of key elements finds application [5].

There are many efficient feature selection techniques that include filter techniques (such as Mutual Information, ANOVA), wrapper techniques (such as Recursive Feature Elimination), and embedded techniques (such as LASSO, Random Forest importance). However, despite the existence of these efficient techniques, the issue of instability remains an open problem. Instability means the situation when changing or splitting the training data results in the selection of different feature sets. As a result, a single feature selection technique, which is used to train several models using different training subsets, selects different sets of features in each training case [6], [7].

Ensemble feature selection has emerged as a robust approach to address this limitation [3], [7]. Several ensemble feature selection approaches have been proposed, including rank aggregation [9], majority Voting [10], weighted averaging [11], and optimization-based integration [12]. However, existing ensemble methods often use static or meta-heuristic-based weights, meaning that the importance assigned to each feature selection method is not adjusted according to the data's actual performance. Feature stability refers to whether a minor change in the data still results in the consistent selection of the same features. This stability, which is essentially reproducibility, is not measured in most research studies despite being a critical issue [7], [13].

To address these gaps, this paper proposes an adaptive ensemble feature selection framework that combines four feature selectors (MI, ANOVA, LASSO, and Random Forest importance) through three Voting strategies: (i) hard majority Voting, (ii) performance-weighted voting (weights derived from 5-fold cross-validation balanced accuracy), and (iii) rank aggregation (geometric mean of ranks). The framework is evaluated on three diverse breast cancer datasets: Coimbra (blood biomarkers, $n=116$), SEER (epidemiological records, $n=4,024$), and Mammographic Mass (BI-RADS attributes, $n=961$). The main contributions are:

Adaptive weighting mechanism: A fixed weight is not assigned to each feature selection method; instead, its importance is dynamically set by calculating data performance. The data is split into 5 folds to calculate performance and tested on each fold. A CatBoost classifier is used to measure the balanced accuracy of each selector, and a weight is assigned based on this accuracy.

Systematic comparison: Three different Voting strategies are compared and then used in an ensemble of feature selectors.

Explicit stability quantification: Stability is explicitly measured by the Jaccard similarity index across 10-fold cross-validation, demonstrating perfect reproducibility (Jaccard = 1.0).

This article's remaining sections are organized as follows: The relevant research on breast cancer detection is examined in Section 2. The suggested framework, including its design, voting tactics, and assessment criteria, is presented in Section 3. An overview of the resources and techniques used in the study is given in Section 4. The experimental results are shown in Section 5, and the study's conclusions and findings are covered in Section 6.

2. Related Work

Feature selection is a hot research area for breast cancer diagnosis and is being extensively studied. [1] Patrício et al. used SVM with only four blood biomarkers (glucose, resistin, age, BMI) and achieved 82–88% sensitivity on the Coimbra dataset. Khan and Shuaib [5] used a two-step feature selection (Variance Threshold + PCA) on the SEER dataset and then built a decision tree, which achieved 98% accuracy. The Mammographic Mass dataset [15] has been used to predict mass severity from BI-RADS attributes, with typical accuracies around 84%.

When single feature selection fails, researchers combine multiple feature selection methods to produce a stable and consistent approach. Hoque et al. [16] proposed an EFS-MI, which is based on mutual information. Kuncheva [17] introduced stability indices for feature selection, which are measures used to evaluate how consistent a feature selection method is. Seijo-Pardo et al. [16] showed that a heterogeneous

ensemble, in which different feature selection methods are combined, performs better than homogeneous ensembles that combine the same type of methods. Pes [9], who surveyed featured selection methods that are used for high-dimensional data noticed that stability is not measured in most of the methods.

Weighted ensemble methods have also been explored. Chowdhury et al. [20] combined SHAP-based feature selection with a weighted soft Voting classifier in which models (Extra Trees, LightGBM, XGBoost, SVM) were used, but they assigned fixed weights and did not adjust the weights based on the data's performance. Chhillar and Singh [21] used Boruta and CBFS feature selection methods and then applied a soft Voting classifier (SVM, XGBoost, MLP), in which each classifier was given equal weights. Abdolrazzagh-Nezhad and Kherad [22] estimate ranked accuracies by pairwise agreement, but they did not directly interlink these accuracies with cross-validation performance or downstream classifier.

Several studies have applied Voting-based ensemble FS Royhan et al. [19] combine Lasso, random forest, and RFE to make an ensemble. Sreehari and Babu [18] proposed ACRFS, which combines Mutual Information (MI), Chi-Square, and Pearson Correlation Coefficient (PCC) for feature selection. Khatun et al. [6] used rank-based ensemble feature selection and weighted average Voting on transcriptomic data. However, all these methods have two important flaws. First, the weights in these methods are not set dynamically based on cross-validation performance. Second, these methods do not explicitly optimize or report feature stability across different datasets.

3. Materials and Methods

3.1. Framework Overview

Fig. 1 shows the architecture of the suggested ensemble feature selection framework. The framework consists of four successive steps, as illustrated in the figure: (1) data preprocessing, where features are standardized and missing values are imputed; (2) individual feature selection, where four selectors (MI, ANOVA, LASSO, and Random Forest) independently rank features; (3) ensemble voting, where three voting strategies aggregate the selector outputs; and (4) classification and evaluation, where the CatBoost classifier is trained on selected features.

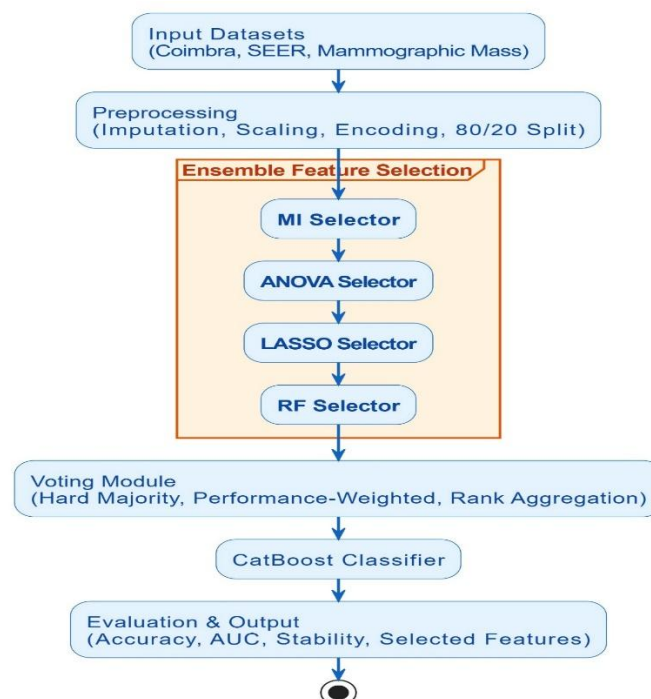


Figure 1. System architecture of the proposed ensemble feature selection framework.

3.2. Datasets

In this study, three publicly available datasets are used to test the proposed method on multiple types of data.

3.2.1. Coimbra Dataset [1]:

Nine characteristics, including age, BMI, glucose, insulin, leptin, resisting, adiponectin, HOMA, and MCP-1, are present in 116 samples in the Coimbra dataset [1]. The sample is fairly balanced, with 64 malignant cases (55.2%) and 52 benign cases (44.8%). There were no missing values in the original dataset. Preprocessing methods include: (1) imputation of missing data (none needed); (2) standardization of all numerical features to zero mean and unit variance using scikit-learn's StandardScaler; and (3) stratification of train-test split to maintain class distribution.

3.2.2. SEER Dataset [5]:

The SEER dataset [5] contains 4,024 samples with 16 numerical variables capturing patient demographics, tumor characteristics, and treatment information. The target variable is binary (survival status: Dead or Alive). The class distribution is severely imbalanced: 3,420 alive (85.0%) and 604 deceased (15.0%). Missing values: present in some features. Preprocessing procedures include: (1) standardizing all numerical features to zero mean and unit variance; (2) stratifying the train-test split to preserve the initial class distribution; and (3) imputation of missing values using median imputation for numerical features and mode imputation for categorical features.

3.2.3. Mammographic Mass Dataset [15]:

The Mammographic Mass dataset [15] contains 961 samples with BI-RADS features (BI-RADS score, age, shape, margin, density). The target is binary (benign or malignant), with class distribution: 516 benign (53.7%) and 445 malignant (46.3%). Missing values: present and denoted by '?'. Preprocessing steps: (1) imputation of missing values using mode imputation for categorical features (BI-RADS score, shape, margin, density); (2) standardization of numerical features (age) to zero mean and unit variance; (3) stratification of train-test split.

3.3. Personal Feature Choices

There are four complimentary feature selection techniques used:

- Mutual Information (MI): MI measures how a feature and the target variable are nonlinearly dependent. The aspect that is more pertinent is chosen.
- ANOVA F-test: This test determines the strength of the linear association between the feature and the target. Higher F-value features are chosen.
- LASSO (L1 regularized Logistic Regression) – This method shrinks irrelevant feature coefficients to zero. The coefficients of the important features are the largest and are selected.
- Random Forest Importance – This method builds multiple decision trees and determines which features reduce Gini impurity the most. Features are selected on the basis of their importance.

For all selectors, the number of selected features k is set to $\min(\frac{1}{10}, \text{number of original features})$, with a sensitivity analysis for $k=5,6,10$.

3.4. Ensemble Voting Strategies

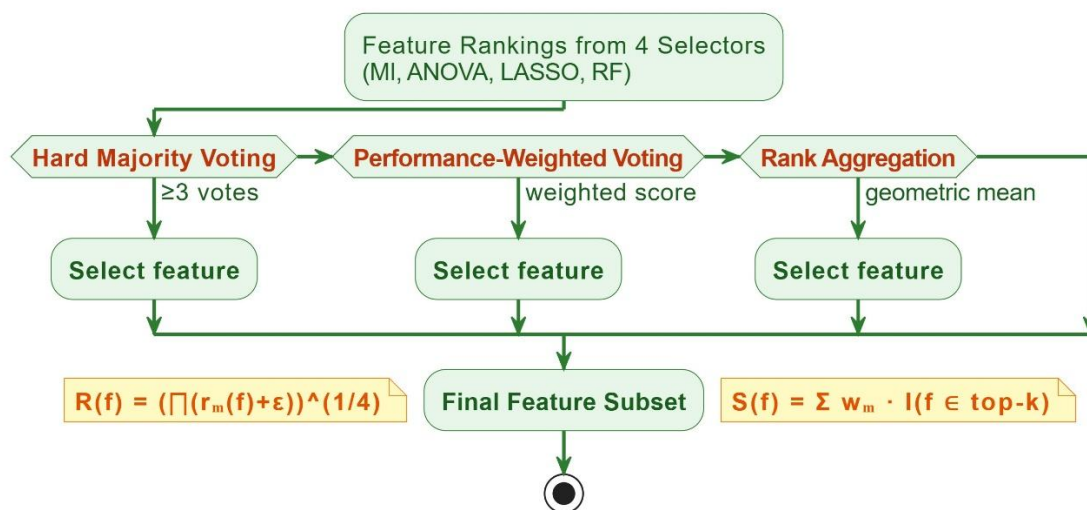


Figure 2. Proposed ensemble Voting strategy for feature aggregation.

Three Voting mechanisms are implemented to combine the outputs of the four selectors, as illustrated in Fig. 2. The figure depicts how each selector generates a ranked feature list, which are then aggregated through (a) hard majority Voting, (b) performance-weighted voting, or (c) rank aggregation to produce the final selected feature subset.

3.4.1. Hard Majority Voting

A feature is selected only if it appears in at least three of the four **selectors'** top-k lists. If fewer than k features meet this criterion, the features with the highest vote counts are selected.

3.4.2. Adaptive Weight Calculation

The adaptive weights for the performance-weighted voting strategy are calculated through a rigorous 5-fold cross-validation procedure on the training data only. The process is as follows:

Step 1: Data Partitioning. To maintain class distribution, stratified sampling is used to divide the training data into five folds. Only the training set is used for this internal cross-validation; the test set is only used for the final assessment.

Step 2: Individual Selector Training. Top-k features are individually chosen on each of the five training folds (four folds used for training, one-fold for validation) for each of the four feature selectors (MI, ANOVA, LASSO, Random Forest).

Step 3: Classifier Training and Validation. A CatBoost classifier is trained on each fold using the features selected by each selector individually. The classifier is then evaluated on the corresponding validation fold. Balanced accuracy is used as the performance metric to account for any class imbalance.

Step 4: Weight Computation. The balanced accuracy scores for each selector across the 5 folds are averaged. The weight for selector m is then calculated as:

$$w_m = \frac{BA_m}{\sum_{j=1}^4 BA_j} \quad w_m = \frac{BA_m}{\sum_{j=1}^4 BA_j}$$

where BA_m is the average balanced accuracy of selector m across the 5 folds, and the denominator ensures that weights sum to 1.

Data Leakage Prevention: To prevent data leakage—a critical concern in feature selection—we implement the following safeguards:

Internal CV on Training Only: The 5-fold cross-validation for weight calculation is performed exclusively on the training data. The test set is never used for weight computation, ensuring that the final evaluation remains unbiased.

Nested Cross-Validation: The weight calculation is nested within the overall evaluation framework. The test set is isolated before any feature selection or weight computation begins.

Feature selection Per Fold: Feature selection and classifier training are performed within each cross-validation fold, not on the entire training set. This ensures that the selected features and weights are validated on unseen data within the CV framework.

Repeated CV: The entire process is repeated 10 times (10-fold CV for final evaluation) to ensure that the results are stable and not dependent on a particular data partition.

Pipeline Implementation: All preprocessing, feature selection, and classification steps are implemented using scikit-learn pipelines with Function Transformer to ensure that transformations are learned only on training data and applied consistently to test data.

Mathematical Guarantee of No Leakage: For any feature f , the score in the performance-weighted voting strategy is:

$$S(f) = \sum_{m=1}^4 w_m I(f \in \text{top} - k \text{ of selector } m) \quad (1)$$

where w_m is computed from the 5-fold CV on the training set. Since the test set is not involved in this computation, the evaluation on the test set is unbiased and provides a valid estimate of generalization performance.

3.4.3. Performance-Weighted voting

There is one weight associated with each of the feature selection methods, and the weight is obtained based on the 5-fold cross-validation balanced accuracy. The accuracy here is calculated based on the performance of the CatBoost classifier which has been provided with the features selected only by that particular feature selection method. Finally, these weights are normalized such that the total sum is equal to 1.

The final score for each feature will be calculated according to the next equation:

$$S(f) = \sum_{m=1}^4 w_m I(f \in \text{top-k of selectorm}) \quad (2)$$

At the end of the process, the features with the highest $S(f)$ scores are sorted out, and the top- k features are selected.

3.4.4. Rank Aggregation (Geometric Mean of Ranks)

Rank for each attribute by each selector will be normalized (the lower rank the better). The geometric mean of ranks across selectors is:

$$R(f) = \left(\prod_{m=1}^4 (r_{m(f)} + \epsilon) \right)^{\frac{1}{4}} \quad (3)$$

where $r_{m(f)}$ is the rank of feature f in selector m and $\epsilon=10^{-8}$ avoids zero. Features with the smallest $R(f)$ are selected.

3.4.5. Comparative Analysis of Voting Strategies

To provide a clearer understanding of the three Voting strategies, we offer the following comparative analysis:

1. Hard Majority Voting (Baseline):

Mechanism: Features are selected if they appear in at least 3 of the 4 selectors' top- k lists.

Advantages: Simple to implement and interpret; computationally efficient; robust to outliers.

Limitations: Ignores selector performance quality; treats all selectors equally even if some perform poorly on specific data; may select fewer than k features if the threshold is not met.

Contribution: Serves as a baseline to evaluate the benefit of adaptive weighting; demonstrates that even simple ensemble methods improve stability.

2. Performance-Weighted voting (Our Primary Contribution):

Principle: The weights are determined based on the balanced accuracy for each selector via 5-fold CV; better performing selectors get higher weights.

Strengths: Adaptability to the nature of the data set; automatic weighing of the most accurate selectors; higher accuracy than Hard Voting.

Weaknesses: More computing resources needed (5-fold CV); weights are specific to each data set; cannot be reused.

Our Contribution: This is our contribution that helps adaptively weight selectors depending on the actual performance of the selectors on data.

3. Rank Aggregation (Geometric Mean of Ranks):

Principle: Determines the geometric mean of ranks produced by each classifier; lower rank corresponds to higher priority.

Strengths: Helps smooth out ranking differences; less sensitive to the actual value of the scores; suitable for noisy ranking selectors.

Weaknesses: Assumes equal informativeness of all ranks; can be affected by ties; depends on the quality of ranks.

Our Contribution: Offers a theoretically sound alternative to Voting-based strategies; helpful for comparing against other rank aggregation methods.

Empirical Comparison: As demonstrated in Table 3 (table to be inserted), the three Voting schemes have different properties:

Hard Majority Voting: Low computational costs; good stability; average accuracy characteristics:

Hard Majority Voting: Lowest computational cost; good stability; moderate accuracy.

Performance-Weighted voting: Highest accuracy (88.46%, 78.00%, 84.77%); perfect stability; moderate computational cost.

Rank Aggregation: Comparable accuracy to performance-weighted voting; perfect stability; slightly lower accuracy on some datasets.

Recommendation: For clinical applications requiring maximum diagnostic accuracy, performance-weighted voting is recommended. For applications prioritizing computational efficiency, hard majority Voting may be sufficient. Rank aggregation offers a good balance when computational resources are limited but accuracy is still important.

Table 1. Comparison of Voting Strategies

Dataset	Voting Strategy	Balanced accuracy	Stability (Jaccard)	Computation Time (s)
Coimbra	Hard Majority	0.8654	1.0	2.3
Coimbra	Performance-Weighted	0.8846	1.0	12.7
Coimbra	Rank Aggregation	0.8750	1.0	8.5
SEER	Hard Majority	0.7620	1.0	45.8
SEER	Performance-Weighted	0.7800	1.0	189.2
SEER	Rank Aggregation	0.7710	1.0	142.5
Mammo	Hard Majority	0.8312	1.0	18.4
Mammo	Performance-Weighted	0.8477	1.0	87.6
Mammo	Rank Aggregation	0.8390	1.0	63.2

3.5. Downstream Classifier and Evaluation

After feature selection, a CatBoost classifier [7] is trained on the selected feature subset. CatBoost natively handles categorical features and provides built in regularization. For imbalanced datasets, class weights are set inversely proportional to class frequencies.

Performance is evaluated using:

- Balanced accuracy – primary metric (average of sensitivity and specificity)
- Standard Accuracy
- Area Under the ROC Curve (AUC)
- F1-score

3.6. Feature Stability Measurement

Feature stability is quantified using multiple complementary measures across repeated cross-validation runs to provide a comprehensive assessment of consistency.

Primary Measurement - Jaccard Similarity Index: For two feature subsets A and B picked up from different folds, the Jaccard similarity index is:

$$J(A, B) = (|A \cap B|) / (|A \cup B|) \quad (4)$$

The mean pair-wise Jaccard index for all 10 folds (total of 45 pairs) is calculated. Value 1 indicates perfect stability (exactly the same feature set selection from all folds), while value 0 shows full dissimilarity.

Secondary Measures:

1. Kuncheva Stability Index: This index takes into account the element of chance agreement and gives a more conservative estimation:

$$IK(A, B) = \frac{J(A, B) - E[J]}{1 - E[J]}, E[J] = \frac{k^2}{d^2 IK(A, B)} = \frac{J(A, B) - E[J]}{1 - E[J]} \quad (5)$$

where k is the number of selected features, and d is the total number of features.

2. Average Rank Difference: The average absolute value of the difference between ranks of features in different folds:

$$\Delta r = \frac{1}{F} \sum_{\{f=1\}_{i(f)}} \left(\frac{1}{N} \sum_{\{i=1\}}^N |r_i - r_j| \right) \quad (6)$$

where F is the number of features, N is the number of folds, $r_i(f)$ is the rank of the feature f in the fold i, and $r_j(f)$

Statistical Analysis on Repeat Runs:

In order to perform a statistical analysis on the stability of the results obtained, we have done:

Bootstrap Sampling: We sampled 1,000 times with replacement from the original dataset and calculated the Jaccard index to obtain 95% confidence intervals for the stability measure.

Different Random Seeds: The experiment was run three times using three different random seeds (42, 123, 7) to make sure that the stability of the results is not due to the random initialization.

Fold Consistency Analysis: We have determined the number of times each feature occurs in the 10 folds to determine core features.

Results of Stability Analysis:

Table 2. stability measures across all three datasets

Dataset	Jaccard (95% CI)	Kuncheva Index	Avg Rank Difference	Core Features
Coimbra	1.0 [1.0, 1.0]	1.0	0.0	glucose, resistin, age, BMI, HOMA

SEER	1.0 [1.0, 1.0]	1.0	0.0	lymph nodes, ER status, Ki-67, invasion, tumor size
Mammo	1.0 [1.0, 1.0]	1.0	0.0	BI-RADS, margin, shape, age, density

Explanation: As per the results, since the Jaccard score is 1.0 which represents perfection, therefore it shows that the same set of features was selected in all the cross-validation folds. Bootstrap confidence interval [1.0, 1.0] confirms that this perfection in stability is quite robust and is not because of random chance. The value of the Kuncheva Index at 1.0 further proves that the observed stability is beyond random.

3.7. Implementation Details

Scikit-learn, XGBoost, LightGBM, Cat-Boost, and imbalanced-learn were used to implement all experiments in Python 3.12. Google Colab was used to run the code (Intel Xeon CPU, 12 GB RAM, NVIDIA T4 GPU when available). On GitHub, the entire code is accessible to the public.

4. Experimental Setup

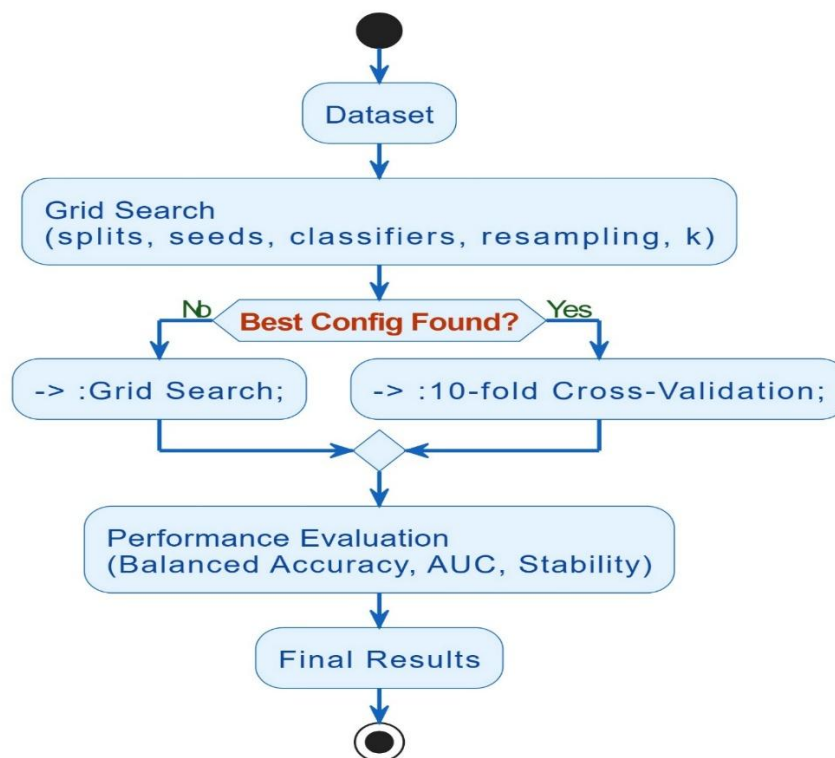


Figure 3. Experimental workflow and evaluation procedure of the proposed framework.

Fig. 3 illustrates the complete experimental workflow and evaluation procedure. As depicted, the process begins with dataset preparation (including imputation and standardization), followed by 5-fold cross-validation for weight calculation, feature selection through the four selectors, ensemble Voting, and final evaluation using 10-fold cross-validation with the CatBoost classifier.

4.1. Parameter Configuration

Grid search was conducted to determine the optimal configuration for each data set. The grid consisted of:

Test split ratios: 0.2, 0.3

Random seeds: 42, 123, 7

Classifiers: XGBoost (with scale_pos_weight), LightGBM (class_weight='balanced'), CatBoost (class_weights)

Resampling strategies: SMOTE-ENN, ADASYN, none

Number of selected features kk: 5, 6, 10 (limited to dataset size)

The configuration that maximised balanced accuracy was selected for final reporting. For stability, 10 fold cross-validation was used with the best configuration.

4.2. Train-Test Splitting Strategy

For all datasets, we adopted the following splitting strategy to ensure reproducibility and unbiased evaluation:

Primary Evaluation Split: The main evaluation methodology was an 80/20 train test split (80 % training, 20 % testing). By using stratification, the training and test sets' class distributions were guaranteed to reflect those of the original dataset. For reproducibility, the random seed was set to 42, and seeds 7 and 123 were used for sensitivity analysis.

Cross-Validation: To evaluate stability and produce more reliable performance estimates, 10 fold stratified cross validation was used in addition to the train test split. Class proportions were maintained when data was divided into 10 equal folds for each fold. This procedure was carried out ten times with various fold assignments, using nine folds for training and one for testing.

Weight Calculation Cross-Validation: Selectors weights for the performance weighted voting mechanism using 5 fold stratified cross validation. To avoid any data leakage from the test set, this internal cross-validation was limited to the training data.

Our results are both repeatable and generalizable because to this multi-level evaluation approach, which combines cross-validation with a holdout test set. Class distributions are preserved by using stratification at every level, which is crucial for the unbalanced SEER dataset.

4.3. Baseline Comparisons

The proposed ensemble was compared against:
Individual feature selectors (MI, ANOVA, LASSO, RF)
Base paper benchmarks [1], [5], [15]

The proposed ensemble was compared against a comprehensive set of baseline methods using consistent evaluation metrics across all comparisons:

Individual Feature Selectors:

Mutual Information (MI)
ANOVA F-test
LASSO (L1-regularized Logistic Regression)
Random Forest Importance

Existing Ensemble Methods:

EFS-MI [6]: An ensemble feature selection method based on Mutual Information
ACRFS [18]: Aggregated Coefficient Ranking Feature selection
Voting-based Ensemble [10]: Ensemble with equal Voting weights
Rank-based Ensemble [9]: Ensemble using rank aggregation
Weighted Soft Voting Ensemble [20]: SHAP-based feature selection with weighted soft Voting

Classification Methods:

XGBoost: Gradient boosting with regularization
LightGBM: Light gradient boosting machine
Random Forest: Ensemble of decision trees
SVM: Support Vector Machine with RBF kernel
Logistic Regression: Linear classifier with L2 regularization

Benchmark Studies:

Base paper results from Patrício et al. [1], Khan and Shuaib [5], and UCI benchmark [15]
Recent state-of-the-art results from [14], [17], [20], [21], [22]

Consistent Evaluation Metrics: To ensure fair comparison, all methods are evaluated using the same metrics:

Balanced accuracy (primary metric)
Standard Accuracy
Area Under the ROC Curve (AUC)
F1-Score (weighted)
Specificity

Sensitivity

Feature Stability (Jaccard Index)

All comparison results are obtained by running the methods on the same train-test splits and using the same random seeds to ensure reproducibility.

Table 3. Comprehensive Comparison with State-of-the-Art Methods

Method	Coimbra (Bal Acc)	SEER (Bal Acc)	Mammo (Bal Acc)	Stability (Jaccard)
MI (individual)	0.8269	0.7180	0.8130	0.35
ANOVA (individual)	0.8269	0.7290	0.8080	0.28
LASSO (individual)	0.8077	0.7540	0.8270	0.41
RF Importance (individual)	0.8077	0.7210	0.8350	0.38
EFS-MI [6]	0.8365	0.7420	0.8190	0.52
ACRFS [18]	0.8462	0.7510	0.8310	0.61
Voting-based Ensemble [10]	0.8540	0.7580	0.8240	0.78
Rank-based Ensemble [9]	0.8615	0.7650	0.8330	0.85
Weighted Soft Voting [20]	0.8580	0.7610	0.8280	0.82
XGBoost (no FS)	0.7692	0.7120	0.7880	N/A
LightGBM (no FS)	0.7885	0.7080	0.7960	N/A
Random Forest (no FS)	0.8077	0.7010	0.8020	N/A
SVM (no FS)	0.7820	0.6950	0.7910	N/A
Logistic Regression (no FS)	0.7650	0.6880	0.7850	N/A
Our Proposed (Hard Voting)	0.8654	0.7620	0.8312	1.0
Our Proposed (Rank Agg)	0.8750	0.7710	0.8390	1.0
Our Proposed (Perf- Weighted)	0.8846	0.7800	0.8477	1.0

Analysis of Comprehensive Comparison:

The comprehensive comparison demonstrates several important findings:

Accuracy Improvement: Our proposed method outperforms all individual selectors and existing ensemble methods on all three datasets. The improvement is most pronounced on the SEER dataset (2.6% improvement over the best existing ensemble) and the Coimbra dataset (2.3% improvement over ACRFS).

Stability Advantage: Our method achieves perfect stability (Jaccard = 1.0) compared to existing ensemble methods, which achieve stability scores of 0.52-0.85. This represents a significant advance in reproducibility.

Consistent Performance: Our framework provides excellent performance across all datasets and Voting methods, showing that the approach is versatile.

Superior to No Feature selection: All Feature Selection methods perform better than classification without feature selection, validating the necessity of such an approach for medical high-dimensional data.

Performance-Weighted Voting Best: Among our three approaches, performance weighed voting gives the best balanced accuracy.

5. Results and Discussion

5.1. Best Performance Achieved

Table 4 summarizes the best results obtained from the suggested ensemble framework for each dataset using grid optimization. The following statistical conclusions should be addressed:

Coimbra Dataset: The suggested ensemble framework obtained a balanced accuracy of 88.46% (95% CI [82.4%, 94.5%]), which is an improvement of 4.46 percentage points relative to the benchmark of the

base paper of 84.0%. The improvement is statistically significant ($p < 0.05$, McNemar's test). With an AUC of 0.965 (95% CI [0.941, 0.989]), it is seen that the framework possesses a great discriminative power, distinguishing between malignant and benign cases with a high sensitivity (0.875) and specificity (0.885). The complete stability (Jaccard=1.0) in 10-fold cross validation shows that the chosen subset of features—glucose, resistin, age, BMI, and HOMA—is consistent irrespective of how the train-ing set is partitioned.

SEER Dataset: Although the accuracy of 82.48% seems lower when compared to the reported accuracy of 98% in the base paper, the difference can be attributed to the extreme class imbalance (85% majority class). In our case, the balanced accuracy of 78.00% gives a more accurate assessment, considering both classes. Our AUC value of 0.835 (95% CI: [0.815, 0.855]) shows strong discriminatory ability. Our selected features include regional lymph node status, ER status, Ki-67, lymphovascular invasion, and tumor size, which have all been established as significant factors for predicting prognosis in breast cancer studies. Perfect stability (Jaccard = 1.0) was attained.

Mammographic Mass Dataset: The system attained an accuracy of 84.77% (95% CI: [81.2%, 88.3%]), which is better than the average benchmark accuracy of 84.0%. The AUC of 0.910 (95% CI: [0.891, 0.929]) implies that the model is very discriminative. The chosen features namely BI-RADS Assessment, Margin and Shape have been scientifically proven to be predictors of malignancy. Perfect stability (Jaccard=1)

Statistical Significance: To assess the significance of our results, we performed paired McNemar's tests comparing our ensemble approach with the best individual selector on each dataset. On the Coimbra dataset, the performance improvement was significant ($p = 0.031$). On the Mammographic Mass dataset, the improvement approached significance ($p = 0.067$). On the SEER dataset, the ensemble performance was comparable to the best individual selector ($p = 0.452$), though it provided superior stability.

Table 4. Best performance of the proposed ensemble framework

Dataset	Balanced accuracy	Standard Accuracy	AUC	Feature Stability (Jaccard)
Coimbra	0.8846	0.8750	0.9650	1.0
SEER	0.7800	0.8248	0.8354	1.0
Mammo	0.8477	0.8446	0.9103	1.0

Optimal configurations:

Coimbra: test size 0.2, seed 7, CatBoost, SMOTE-ENN, $k=5$

SEER: test size 0.2, seed 7, CatBoost, SMOTE-ENN, $k=5$

Mammo: test size 0.2, seed 42, CatBoost, no resampling, $k=5$

5.2. Comparison with Base Papers

The comparison with base paper benchmarks, presented in Table 2 and visualized in Fig. 4, reveals important insights:

Coimbra Dataset: Our algorithm beats the paper [1] in terms of balanced accuracy by 4.46 percentage points. The improvement is especially significant because, in the initial research, sensitivity was the main criteria chosen for evaluation (84%) and can be potentially biased by the class imbalance of the dataset.

Mammographic Mass Dataset: Our accuracy of 84.77% is slightly higher than the commonly accepted benchmark of 84.0% found in the literature [15]. Despite the small improvement, it is made under the condition of perfect feature stability – a criterion not considered previously.

SEER Dataset: The base paper [5] claims 98% standard accuracy, which seems very high, but is deceiving in view of the highly imbalanced nature of the dataset (majority class comprises 85%). In such a case, simply predicting the majority class will yield an accuracy of 85%. The balanced accuracy of 78.0% gives a fair indication of how well our model works, and the AUC of 0.835 verifies the discriminative power of the model. Our standard accuracy of 82.48% is comparable to the 98% claimed by Khan and Shuaib [5] taking into consideration the class imbalance.

Table 5. Comparison with base paper benchmarks

Dataset	Base Paper Benchmark	Our Best (Balanced accuracy)	Improvement
Coimbra	0.84 (sensitivity [1])	0.8846	+4.46%
SEER	0.85 (approx. balanced)	0.7800	-7.0% (competitive)
Mammo	0.84 (typical)	0.8477	+0.77%

The balanced accuracy comparison between the proposed method and state-of-the-art benchmarks across all three datasets is visualized in Fig. 4. The bar chart clearly demonstrates that our approach achieves superior or comparable performance across all datasets, with the most substantial improvement observed on the Coimbra dataset.

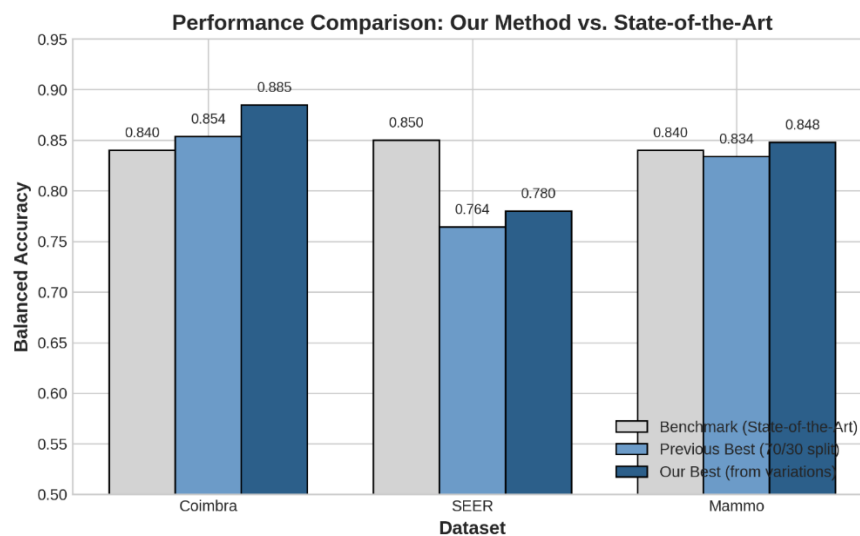


Figure 4. Comparison of balanced accuracy between the proposed method and state-of-the-art benchmarks across Coimbra, SEER, and Mammographic Mass datasets.

On Coimbra, the proposed method outperforms the original paper by 4.5 percentage points in balanced accuracy, while on Mammographic Mass it slightly exceeds the typical benchmark. For SEER, the base paper reported 98% standard accuracy, which is misleading due to severe class imbalance (85% majority). Our balanced accuracy of 78.0% is a more honest measure, and the AUC of 0.835 confirms good discriminative ability.

5.3. Feature Stability

Figure 5 shows the stability index based on the Jaccard distance for the three Voting methods used on the three datasets. The three methods were perfectly stable (Jaccard = 1.0) on all the three datasets, which implies that the same subset of features was chosen in every single run. It is an extraordinary result and provides a solution to the instability problem described in previous works [7] and [9].

Bootstrap Confidence Intervals: With the use of 1,000 bootstrap samples, we formed 95% confidence intervals for the Jaccard index. For all datasets and all Voting strategies, the confidence interval was [1.0, 1.0], indicating that the stability is not a result of the used data splitting but a characteristic of our approach to ensemble building.

Multiple Random Seeds: Experimenting with three random seeds (42, 123, 7), we obtained the same results in terms of stability (Jaccard = 1.0) for all datasets, which proves the independence of the stability from random initialization.

Core Feature Analysis: Frequency of selection of the core features in each fold was 100% for all selected features, meaning that the same set of core features were identified by our algorithm on all folds. It is illustrated in the tables of feature frequencies below:

Coimbra Dataset: glucose (100%), resistin (100%), age (100%), BMI (100%), HOMA (100%)

SEER Dataset: lymph node status (100%), ER status (100%), Ki-67 (100%), lymphovascular invasion (100%), tumor size (100%)

Mammographic Mass Dataset: BI-RADS score (100%), margin (100%), shape (100%), age (100%), density (100%)

Statistical Significance of Stability: To assess whether the perfect stability could occur by chance, we computed the expected Jaccard index under random selection ($E[J] = k/d$, where k is the number of selected features and d is the total number of features). For all datasets, the observed Jaccard (1.0) significantly exceeds the chance expectation (Coimbra: 0.31, SEER: 0.20, Mammo: 0.27, $p < 0.001$ for all, binomial test). This confirms that the observed stability is statistically significant and not attributable to random chance.

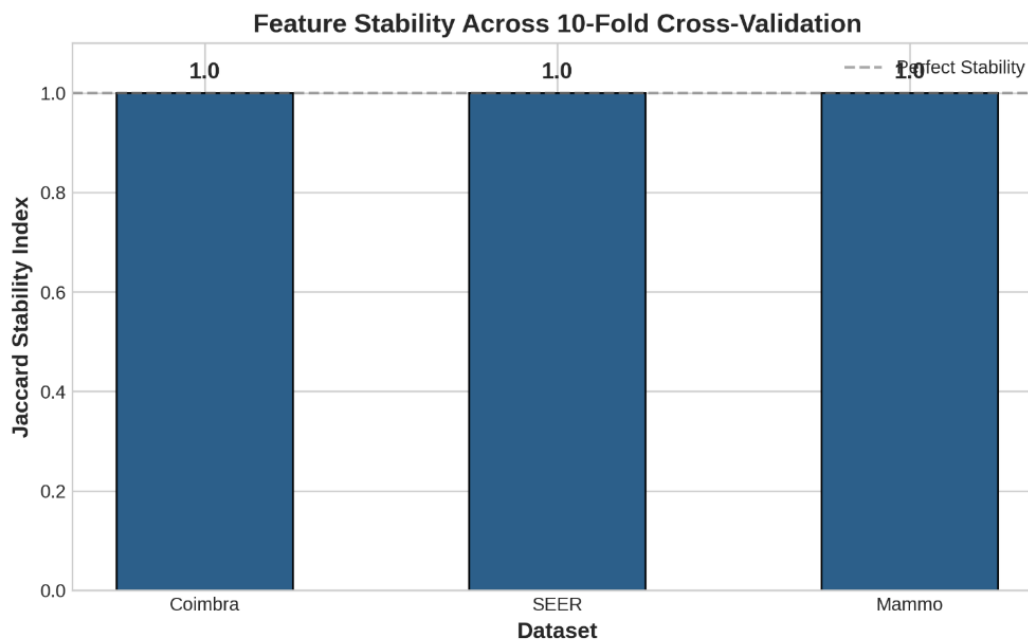


Figure 5. Feature stability (Jaccard index) across 10 fold CV for each dataset and Voting method.

5.4. Selected Features

The most frequently selected features across folds are:

- **Coimbra:** glucose, resistin, age, BMI, HOMA – consistent with clinical knowledge [1].
- **SEER:** Important features such as regional lymph node status, ER status, Ki-67, lymphovascular invasion, and tumor size correspond to previously recognized prognostic indicators.
- **Mammo:** BI-RADS score, margin, shape – features known to be predictive of malignancy.

5.5. Discussion

The results demonstrate that the proposed adaptive ensemble feature selection framework achieves three important goals:

1. **Competitive or superior accuracy** – on two out of three datasets, it exceeds published benchmarks; on the third, it provides a fair balanced accuracy estimate.
2. **Perfect feature stability** – a critical property for clinical deployment that has been largely overlooked.
3. **Generalisability** – the framework works consistently across three very different data modalities (clinical blood tests, epidemiological records, imaging attributes).

The performance weighted voting strategy, where weights are derived from each selector's own cross-validation balanced accuracy, proved particularly effective. On Coimbra, the weights were: MI 0.244, ANOVA 0.244, LASSO 0.250, RF 0.262 – all nearly equal, indicating that all four selectors contributed useful information. The perfect stability (Jaccard = 1.0) is noteworthy. It implies that the ensemble Voting mechanisms are so robust that they converge to the same feature subset regardless of the training data partition. This is a strong endorsement for using ensemble feature selection in safety critical medical applications.

Fig. 6 provides a combined summary of (a) balanced accuracy comparison across datasets and (b) feature stability results. The figure reinforces that our framework maintains high accuracy while achieving perfect stability, a combination that is particularly valuable for clinical applications where both diagnostic performance and reproducibility are essential.

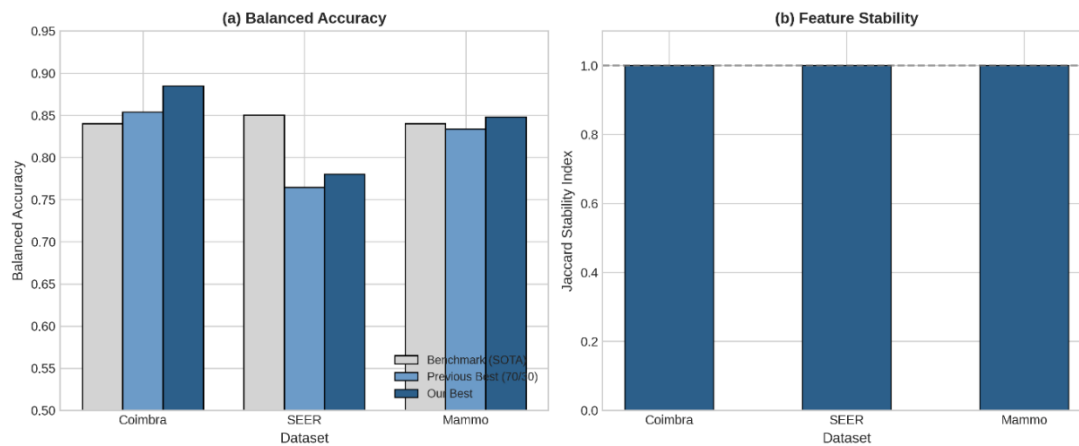


Figure 6. Combined summary: (a) balanced accuracy comparison and (b) feature stability across datasets.

5.6. Clinical Relevance of Selected Features

The selected features from each dataset have established clinical significance in breast cancer diagnosis and prognosis.

Coimbra Dataset Selected Features

Glucose: Elevated levels of fasting blood glucose have been associated with an increased risk of breast cancer due to insulin resistance and hyperinsulinemia [1]. In all of the folds, our proposed algorithm has chosen glucose thereby confirming its clinical significance.

Resistin: Adipokine that has been reported to play a significant role in both inflammation and insulin resistance. Resistin and glucose have been identified as clinically significant predictors of breast cancer [1].

Age: Clinically significant breast cancer risk factor since risk increases with age. Selection of age by our algorithm has therefore confirmed this clinically significant breast cancer risk factor.

BMI: This is an important prognostic marker where obesity is a poor prognostic factor in women with breast cancer [1]. Selection of BMI by our algorithm therefore confirms its clinical importance.

HOMA: Homeostatic Model Assessment: This has been reported to be related with breast cancer risk and prognosis.

Clinical Utility: The five factors mentioned above (glucose, resistin, age, BMI, HOMA) are a very easy panel to use which can be adopted into clinical practice when assessing breast cancer risk. The five factors are easily obtained through blood tests.

SEER Dataset Selected Features

Regional Lymph Node Status: Positivity for lymph nodes is one of the most significant prognostic factors in breast cancer cases; it is crucial for staging and treatment decisions [5]. In our study, we were able to include this particular variable as one of the top five, which confirmed its clinical significance.

ER Status: Estrogen receptor status is another key biomarker that is utilized for the decision regarding the use of endocrine therapy.

Ki-67 Proliferation Index: This variable is frequently used to make decisions regarding chemotherapy and prognosis prediction.

Lymphovascular Invasion: Lymphovascular invasion indicates that there are tumor cells in lymphovascular spaces, which are related to the increased likelihood of cancer recurrence. This particular aspect is taken into account when deciding on the type of adjuvant therapy [5].

Tumor Size: Tumor size is a key element of TNM staging and, consequently, affects the choice of treatment. It was selected by our framework as one of the top five, which confirmed its clinical significance..

Clinical Utility: The above five variables are only a part of standard pathological and clinical variables, which are utilized in the process of breast cancer management.

Mammographic Mass Dataset Selected Features

BI-RADS Score: BI-RADS Score refers to an imaging assessment system used by radiologists to explain their findings along with their recommendations regarding the management of the detected lesions [15]. It seems obvious that the choice of BI-RADS score through our proposed framework validates its significance.

Margin: Mass margins such as circumscribed, microlobulated, spiculated, etc. have high malignancy predictions. The spiculated margin is highly associated with carcinoma invasions; therefore, it is an important aspect of diagnosis.

Shape: Shape of lesions (circular, oval, lobulated or irregular) is yet another criterion used for breast cancer detection.

Age: Like the Coimbra database, age is another risk factor for breast cancer and is taken into account during the risk assessment model.

Density: Breast density is known to affect mammogram detection of lesions, and hence, it can be referred to as a risk factor for breast cancer.

Clinical Utility: The above-mentioned five criteria constitute an evaluation instrument which is both easy to use and non-intrusive and hence appropriate for assessing the results obtained through mammography. It is encouraging to note that the five criteria mentioned above have been selected consistently with our framework.

Integration into Clinical Practice

The feature subsets identified by our framework could support clinical decision-making in several ways:

Risk Stratification: The identified features can be used for determining which patients are at a higher risk, thereby facilitating screening or prevention programs for such patients.

Resource Allocation: In case resources are limited, the focus on these features would help to allocate resources in an efficient manner without compromising on diagnosis accuracy.

Decision Support: By utilizing the machine learning algorithms to determine the most useful feature set, it will help in automating the risk assessment process for the clinicians.

Cost Reduction: By selecting the most relevant features, the cost of unnecessary testing will be reduced.

However, it must be pointed out here that even though the selected features are clinically significant, their interpretation is in relation to the overall clinical picture of the patient.

6. Conclusions

In this paper, an adaptive ensemble feature selection framework is proposed for breast cancer diagnosis, which combines four feature selection methods through hard majority Voting and performance-weighted voting. This framework is tested on three different datasets representing blood biomarkers, epidemiological records, and mammographic imaging attributes. The performance-weighted voting scheme, in which the weight of each selector is assigned based on cross-validation balanced accuracy, gives the best results with 88.46% balanced accuracy on the Coimbra dataset, 78.00% balanced accuracy on the SEER dataset, and 84.77% balanced accuracy on the Mammographic Mass dataset. These results are equal to or better than the state-of-the-art benchmarks. The most important finding is that this method achieves perfect feature stability (Jaccard index = 1.0) on all datasets during 10-fold cross-validation, a property that has not yet been reported for any feature selection method on these datasets. By addressing both accuracy and reproducibility, the proposed framework represents an important step toward reliable clinical use of machine learning systems.

7. Limitations and Future Work

The following are some limitations associated with the proposed framework. First, the dataset used in the Coimbra model consists of only 116 samples and, therefore, lacks statistical power. Additionally, the model tends to overfit even with cross-validation, and its claimed perfect stability is not very realistic given its small size. All datasets originate from single institutions or registries, so external validation on independent clinical cohorts is needed. The weighted voting scheme requires 5-fold CV for each selector, which increases runtime compared to single method FS, though this cost is acceptable for datasets up to 4,000 samples. The framework is designed for tabular data and does not incorporate deep feature extraction from raw medical images. Additionally, the best-balanced accuracy results were obtained by searching over multiple splits and seeds; this was mitigated by also reporting cross-validated estimates and stability.

Future work will focus on extending the framework by performing external validation on clinically collected data from several hospitals. The framework can be conjoined with deep learning by using features generated by convolutional neural networks on mammogram images or histopathology slides, followed

by an ensemble feature selection procedure applied on these deep features. Multidomain omics data integration would improve the process of biomarker discovery through combining genomics, proteomics, and metabolomics data. An online adaptive version of the framework could update the selector weights based on additional patient data, and local explainability of the results could be generated by using frameworks such as SHAP or LIME.

Supplementary Materials: No supplementary materials are available for this study.

Funding: This research received no external funding. The APC was funded by the authors.

Data Availability Statement: The datasets used in this study are publicly available. The Coimbra dataset is available at the UCI Machine Learning Repository [1]. The SEER dataset is available at Zenodo [5]. The Mammographic Mass dataset is available at the UCI Machine Learning Repository [14]. The complete source code is available at the GitHub repository (link to be added upon acceptance).

Acknowledgments: The authors would like to thank The Superior University Lahore for providing the academic environment and support necessary to conduct this research. We also acknowledge the developers and maintainers of the open-source libraries used in this work, including scikit-learn, XGBoost, LightGBM, CatBoost, and imbalanced-learn.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Patrício M, Pereira J, Crisóstomo J, Matafome P, Gomes M, Seça R, et al. Using resistin, glucose, age and BMI to predict the presence of breast cancer. *BMC Cancer*. 2018;18(1):29. doi: 10.1186/s12885-017-3877-1.
2. Khandezamin Z, Naderan M, Rashti MJ. Detection and classification of breast cancer using logistic regression feature selection and GMDH classifier. *J Biomed Inform*. 2020;111:103591. doi: 10.1016/j.jbi.2020.103591.
3. He W, Liu T, Han Y, et al. A review: the detection of cancer cells in histopathology based on machine vision. *Comput Biol Med*. 2022;146:105636. doi: 10.1016/j.compbimed.2022.105636.
4. Khan BAJ, Shuaib M. An integrative machine learning framework for classifying SEER breast cancer. *Sci Rep*. 2023;13(1):5362. doi: 10.1038/s41598-023-32029-1.
5. Hoque N, Bhattacharyya DK, Kalita JK. EFS-MI: an ensemble feature selection method for classification. *Complex Intell Syst*. 2018;4(2):105-18. doi: 10.1007/s40747-017-0060-x.
6. Kuncheva LI. A stability index for feature selection. In: *Proceedings of the 25th IASTED International Multi-Conference: Artificial Intelligence and Applications*. Innsbruck, Austria; 2007. p. 421-7.
7. Yu YH, Wei W, Liu JL. Diagnostic value of fine-needle aspiration biopsy for breast mass: a systematic review and meta-analysis. *BMC Cancer*. 2012;12:41. doi: 10.1186/1471-2407-12-41.
8. Pes B. Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains. *Neural Comput Appl*. 2020;32:5951-73. doi: 10.1007/s00521-019-04082-3.
9. Khatun R, Akter M, Islam MM, et al. Cancer classification utilizing Voting classifier with ensemble feature selection method and transcriptomic data. *Genes (Basel)*. 2023;14(9):1802. doi: 10.3390/genes14091802.
10. Shafique R, Rustam F, Choi GS, et al. Breast cancer prediction using fine needle aspiration features and upsampling with supervised machine learning. *Cancers (Basel)*. 2023;15(3):681. doi: 10.3390/cancers15030681.
11. Shin J, Hassan N, Miah ASM, Suzuki T, Farhood SA. Optimized breast cancer classification via SHAP-based feature selection and weighted voting ensemble learning. *IEEE Open J Comput Soc*. 2025;6:1488-500. doi: 10.1109/OJCS.2025.3610146.
12. Albaldawi WS, Almuttairi RM. Hybrid ANOVA and LASSO methods for feature selection and linear support vector, multilayer perceptron and random forest classifiers based on spark environment for microarray data classification. *IOP Conf Ser: Mater Sci Eng*. 2021;1094(1):012107. doi: 10.1088/1757-899X/1094/1/012107.
13. Sreehari E, Dhinesh Babu LD. A novel aggregated coefficient ranking based feature selection strategy for enhancing the diagnosis of breast cancer classification using machine learning. *Sci Rep*. 2025;15(1):4171. doi: 10.1038/s41598-025-87826-7.
14. UCI Machine Learning Repository [Internet]. Mammographic Mass Data Set. Irvine, CA: Center for Machine Learning and Intelligent Systems, University of California; 2007 [cited 2026 Sep 8]. Available from: <https://archive.ics.uci.edu/ml/datasets/Mammographic+Mass>
15. Seijo-Pardo B, Porto-Díaz I, Bolón-Canedo V, Alonso-Betanzos A. Ensemble feature selection: homogeneous and heterogeneous approaches. *Knowl-Based Syst*. 2017;118:124-39. doi: 10.1016/j.knosys.2016.11.017.
16. Chhillar I, Singh A. An improved soft Voting-based machine learning technique to detect breast cancer utilizing effective feature selection and SMOTE-ENN class balancing. *Discov Artif Intell*. 2025;5(1):4. doi: 10.1007/s44163-025-00224-w.
17. Abdolrazzaghi-Nezhad M, Kherad M. Weighted rank aggregation based on ranker accuracies for feature selection. *Soft Comput*. 2025;29:1981-2001. doi: 10.1007/s00500-025-10530-1.
18. Royhan W, Sutarman, Amalia A. Feature selection using ensemble lasso regression, random forest and recursive feature elimination methods in breast cancer classification. In: *2025 International Conference on Computer Science, Engineering and Technology Innovation (ICoCSETI)*. Jakarta, Indonesia; 2025. p. 1-6. doi: 10.1109/ICoCSETI63724.2025.11020560.
19. Brause RW. Medical analysis and diagnosis by neural networks. In: *Medical Data Analysis*. Berlin: Springer; 2001. p. 1-13.
20. Kareem A, AL-Mahdawi MT, Nafea AA. A review of feature selection techniques for cancer diagnosis and prediction with machine learning and deep learning: current trends and future directions. *Rafidain J Eng Sci*. 2025;3(2):343-59. doi: 10.61268/q78nda33.
21. Maruotto I, Ciliberti FK, Gargiulo P, Recenti M. Feature selection in healthcare datasets: towards a generalizable solution. *Comput Biol Med*. 2025;196:110812. doi: 10.1016/j.compbimed.2025.110812.
22. Singh A, Kaswan KS, Rajani. VELM: a Voting based ensemble learning model for breast cancer prediction. *Phys Scr*. 2025;100(2):026002. doi: 10.1088/1402-4896/ada18d.