



International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

A Hybrid Feature Selection and Soft Voting Ensemble Framework for Accurate Breast Cancer Classification

Manmohan Sahoo¹, Dr. Aswini Kumar Mohanty²

¹Ph.D. Scholar, Biju Patnaik University of Technology (BPUT), Odisha, India; Assistant Professor, Department of Computer Science and Engineering, KMBB College of Engineering & Technology, Bhubaneswar, Odisha, India. E-mail: manmohansahoo@kmbb.in

²Principal, Department of Computer Science and Engineering, Synergy Institute of Technology, Bhubaneswar, Odisha, India. E-mail: asw_moh@yahoo.com

Abstract

Globally, breast cancer remains a major contributor to mortality among women. This disease can be curable if an early and accurate diagnosis is made. In order to improve the accuracy of predictions, this study suggests a hybrid architecture that combines ensemble classification and optimization-based feature selection. This study employs selection algorithms such as Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), and Genetic Algorithm (GA) for selecting the most significant variables in the Wisconsin Breast Cancer Dataset (WBCD). Two strategies are then applied to combine the selected features: (1) a majority voting approach, where features chosen by at least two of the three optimizers are retained, and (2) a union strategy that aggregates all features selected by any of the methods. After feature selection, to get the result we have applied classifiers such as Random Forest, XGBoost, Logistic Regression, and a soft voting-based. Our method achieved 98.68% accuracy, with strong F1-score and ROC-AUC values. These results suggest that the ensemble feature selection and classification strategy significantly improves performance compared to individual techniques, making it an encouraging approach for dependable breast cancer diagnosis.

1 Introduction

Despite advancements in medical care, breast cancer continues to be one of the most prevalent malignancies in women worldwide and a major cause of mortality. Finding breast cancer early and correctly helps a lot in treating it well and saving lives. Clinical test data and medical images, when used properly, can help in early detection. However, the large amount of data used in breast cancer studies can be a problem for computer programs, which may not work well and can become slow.

Feature selection methods are useful in removing useless or repeated data, which helps in better performance of the programs. Advanced optimization methods, including Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), and Genetic Algorithm (GA), have been effectively applied for selecting the most significant data features. These methods copy how nature works to search through lots of data quickly and are known for solving hard problems well.

Even though most studies use only one of these methods using just one way not find all the important features. To fix this our project introduces a new mixed method that uses the good parts of all three optimizers. We used Aco, Pso, and GA on the Wisconsin Breast Cancer Dataset (WBCD) and tried two ways to combine the features they picked.

i) Union strategy, where we keep all features chosen by methods, and
ii) Majority voting strategy, where we keep only the features picked by at least two of the methods. After choosing the best features, we used different machine learning models like Random Forest, XGBoost, Logistic Regression, and a Soft Voting Classifier to check how will the features work in predicting cancer. We are trying to build a strong and correct model for breast cancer detection that uses the power of many optimizers and classifiers.

Our test results show that our combined feature selection method, especially with a group of classifiers, gives better results in terms of accuracy, F1-score, and Roc-Auc. So our method is a good step towards better and faster breast cancer detection using machine learning.

2 Literature Review

Several studies have aimed to enhance breast cancer diagnostic accuracy by applying data mining and machine learning techniques. Early approaches primarily relied on traditional statistical tools or basic classifiers, which often struggled to manage the complexity and high dimensionality of medical datasets.

Using the Wisconsin Breast Cancer Dataset (WBCD), Chaurasia and Pal [1] implemented Decision Trees and Naïve Bayes, obtaining moderate accuracy and highlighting the necessity for improved feature reduction methods. Neural networks have been explored as an alternative for improved performance. Bharati et al. [2] examined different artificial neural network (ANN) models for identifying breast cancer and found that, although they frequently provide high accuracy, the lack of transparency in their internal workings makes them difficult to interpret. Similarly, Karthik et al. [3] used deep neural networks on the WBCD dataset to achieve 98.6% accuracy, but interpretability remained a challenge.

Hybrid and ensemble approaches have since emerged as more effective. Han and Yin [4] developed a hybrid approach for breast cancer classification that used meta-learning in combination with artificial neural networks. By integrating predictions from several base models into a single ANN, their method attained 98.74% accuracy and an F1-score of 98.02%. However, similar to other feature-reduction strategies like PCA, it may still reduce interpretability of the resulting model.

Metaheuristic algorithms have proven effective in feature selection for breast cancer classification. Al-Saedi et al. [5] applied Particle Swarm Optimization (PSO) for selecting relevant medical features, leading to significant performance improvements. Ant Colony Optimization (ACO) and Genetic Algorithms (GA) have similarly demonstrated the ability to explore complex feature spaces effectively [6,7]. More recent work by Yang et al. [8] introduced a hybrid method combining Information Gain with grouped PSO, producing superior results on cancer datasets.

Ensemble learning models such as Random Forest and XGBoost have consistently outperformed individual classifiers in medical diagnostics due to their robustness and generalization capabilities. Kourou et al. [9] showed that ensemble-based approaches improve predictive performance when paired with optimized feature subsets.

Few studies have combined multiple metaheuristic algorithms with ensemble classifiers. O. Fallahzadeh and Z. Dehghani-Bidgoli [6] used ACO to select critical Raman spectral features, achieving 87.7% accuracy with a reduced feature set. X. Deng [10] developed a hybrid XGBoost-GA approach for gene subset selection, achieving strong multi-objective optimization results. B. Sowon et al. [11] employed a Hybrid Particle Swarm Optimization (PSO)-based feature selection approach for high-dimensional breast cancer datasets, demonstrating substantial improvements in classification accuracy over conventional techniques.

Previously study [12] Demonstrated that ACO-enhanced Random Forest models could achieve up to 99.02% accuracy on the WBCD dataset. Karimi et al. [7] introduced new heuristic-driven selection strategies that outperformed traditional optimization techniques. Ameta [13] emphasized the reliability of voting-based ensemble classifiers for effective cancer grading.

The survey of related works points to a clear shift toward merging sophisticated feature selection strategies with ensemble classification models to achieve higher accuracy, better interpretability, and greater robustness in detecting breast cancer, which underpins the methodology used in this research.

3 Methodology

his learning utilizes the Wisconsin Breast Cancer Dataset (WBCD), consisting of 569 records and 32 attributes, as the primary dataset to develop a robust and accurate classification model through ensemble-based feature selection and classifier fusion techniques. The methodology involves the following sequential steps:

3.1 Data Preprocessing

Cleaning and preprocessing are necessary steps to eliminate missing values and noise from dataset. The target variable diagnosis (M for malignant and B for benign) is encoded numerically using label encoding for compatibility with machine learning algorithms. Prior to model training, all features are normalized to maintain uniform contribution across attributes.

3.2 Feature Selection using Optimization Algorithms

To reduce dimensionality and enhance classification performance, three metaheuristic optimization algorithms are applied independently: Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), Genetic Algorithm (GA)

Each algorithm is executed to identify the most relevant subset of features from the dataset that contributes significantly to classification accuracy. The objective function of each optimizer is designed to maximize classification accuracy using a Random Forest classifier during the search process.

3.3 Feature Merging Strategies

Two strategies are adopted to combine the selected features:

- **Union Strategy:** All unique features selected by ACO, PSO, and GA are combined to form a superset.
 - **Voting Strategy:** A feature is selected only if it appears in at least two out of the three optimization outputs (majority voting).
- These strategies aim to preserve feature diversity while maintaining high relevance for classification.

3.4 Classification Algorithms

the refined feature subsets obtained from each method (ACO, PSO, GA, Union, Voting) are evaluated using the following classifiers:

- **Random Forest**
- **XGBoost**
- **Logistic Regression**

Voting Classifier (Soft Voting)

Classes with stronger consensus across models are given a higher probability by the Voting Classifier, which aggregates the predictions from the three base classifiers.

3.5 Performance Evaluation Metrics

We trained and tested the models using a 60/40 stratified split between training and testing data. To evaluate how the models performed, we used the following metrics:

- **Accuracy**
- **Precision**
- **Recall**
- **F1-Score**
- **ROC-AUC Score**

Additionally, ROC curves and bar charts are plotted for visual comparison of different models and feature selection strategies.

Algorithm 1: Ensemble Feature Selection and Classification Framework

Input: WBCD Dataset D with features F and target variable Y

Output: The accuracy, precision, recall, F1-score, and ROC-AUC metric are used to evaluate the model's performance.

Step 1: Data Preprocessing

- Encode target variable Y (Malignant/Benign to 1/0)
- Normalize feature set F
- Split the dataset D into 60% for training and 40% for testing

Step 2: Feature Selection using Optimization Algorithms

for optimizer in [ACO, PSO, GA]:

- Run optimizer to select optimal feature subset $F_{\text{optimizer}} \subseteq F$
- Save $F_{\text{optimizer}}$

Step 3: Merge Feature Subsets

- $\text{Union_Features} \leftarrow F_{\text{ACO}} \cup F_{\text{PSO}} \cup F_{\text{GA}}$
- $\text{Voted_Features} \leftarrow$ features selected by at least 2 of $[F_{\text{ACO}}, F_{\text{PSO}}, F_{\text{GA}}]$

Step 4: Model Training and Evaluation

for feature_set in $[F_{\text{ACO}}, F_{\text{PSO}}, F_{\text{GA}}, \text{Union_Features}, \text{Voted_Features}]$:

for classifier in [Random Forest, XGBoost, Logistic Regression]:

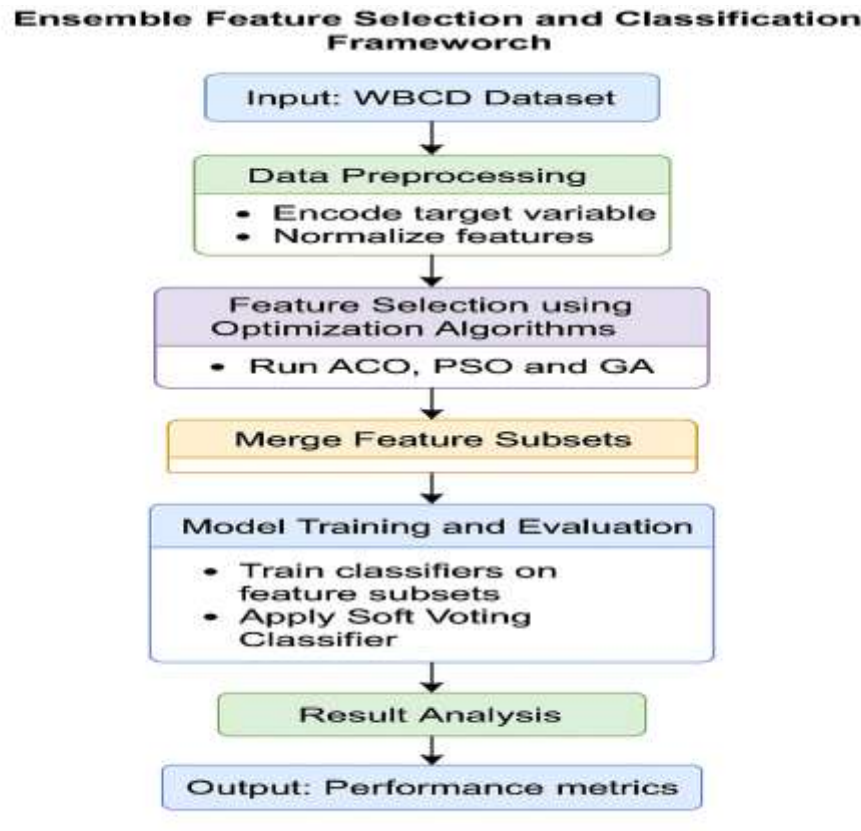
- Train classifier on training data using feature_set
- Evaluate on test data and compute Accuracy, Precision, Recall, F1, ROC-AUC
- Apply Soft Voting Classifier (ensemble of the 3 models)
- Compute ensemble performance metrics

Step 5: Result Analysis

- Plot ROC Curves for each classifier-feature_set combination
- Produce bar charts showing the comparison of accuracy, precision, recall, and F1-score.
- Identify best-performing combinations

Return: All performance results and visualizations

FLOW CHART:



4 Experimental Results

The proposed ensemble-based framework for breast cancer classification was tested through comprehensive experiments using the Wisconsin Breast Cancer Dataset (WBCD). Multiple feature selection methods—Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), Genetic Algorithm (GA), their Union, and Majority Voting (VOTING)—were applied. These selected features were then classified using four algorithms: Random Forest, XGBoost, Logistic Regression, and a Voting Classifier with soft voting.

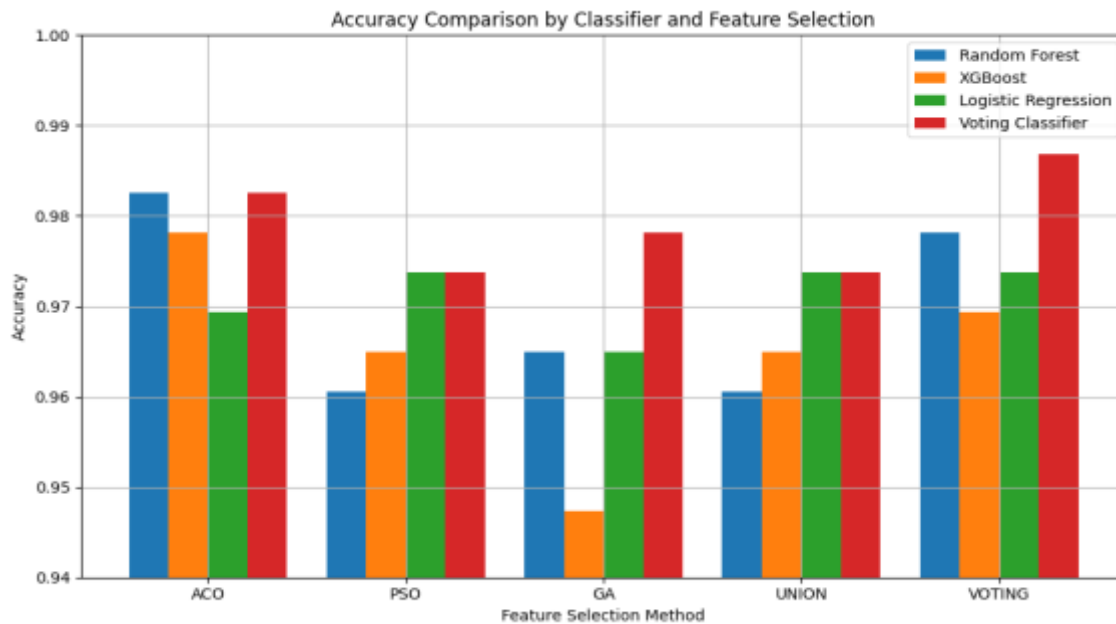
4.1 Evaluation Metrics

We evaluated the models using standard measures such as accuracy, precision, recall, F1-score, and the ROC-AUC (Receiver Operating Characteristic – Area Under the Curve). These metrics were computed on a separate test set created through a 60:40 train–test division. The results indicate the effectiveness of each classifier when applied to the different selected feature sets.

4.2 Accuracy Comparison

The accuracy of all classifiers across various feature selection methods is shown in **Figure 1**. The **Voting Classifier** using **VOTING-based features** achieved the **highest accuracy of 98.68%**, followed closely by XGBoost and Random Forest under ACO and GA-based feature subsets.

Fig. 1 Accuracy comparison across classifiers using different feature selection techniques.



4.3 Recall and F1-Score Analysis

Recall is an essential measure in medical diagnostics since it helps reduce the number of false negatives, and its results are illustrated in **Figure 2**. The Voting Classifier using ACO and Voting-based features achieved nearly perfect recall values.

F1-Score, which balances precision and recall, is shown in **Figure 2.1**. The trends closely mirror those of recall, confirming consistent performance.

Fig. 2 Comparison of recall scores among different classifiers and feature selection techniques.

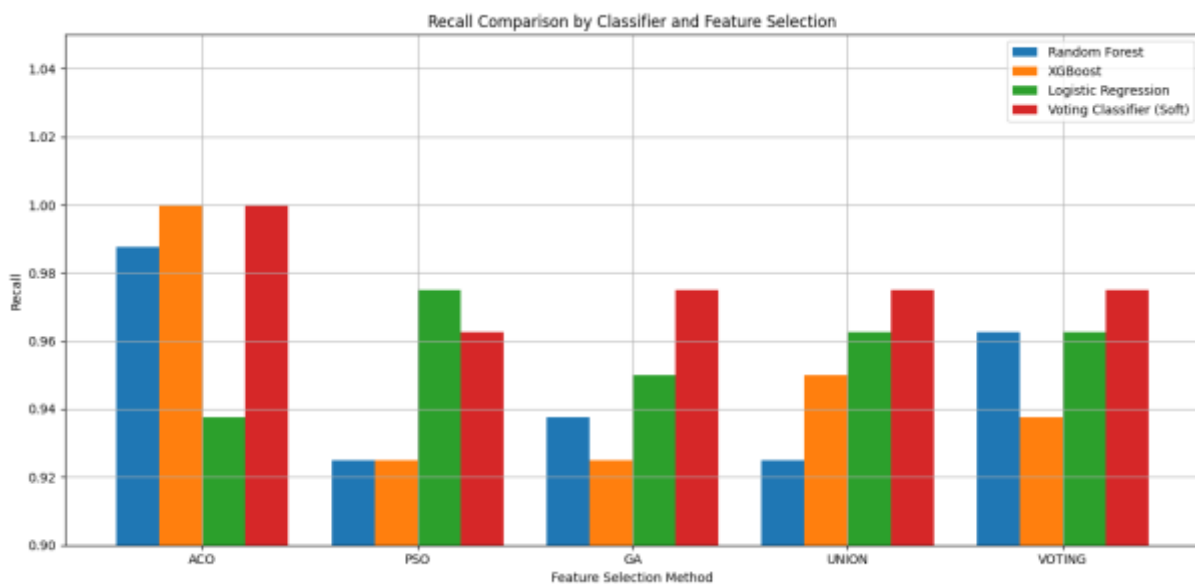
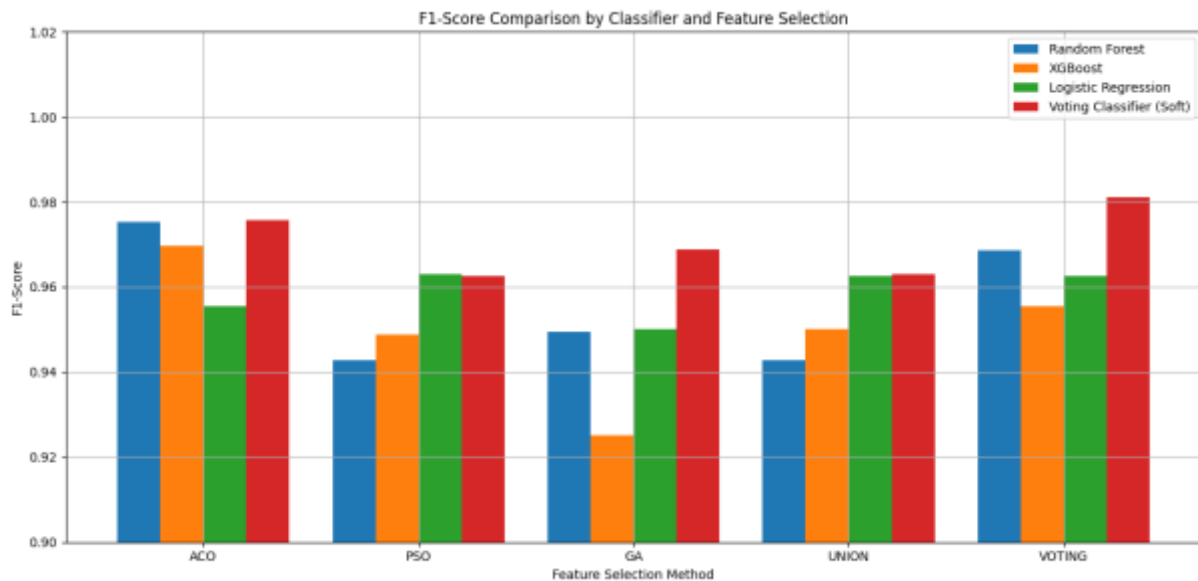
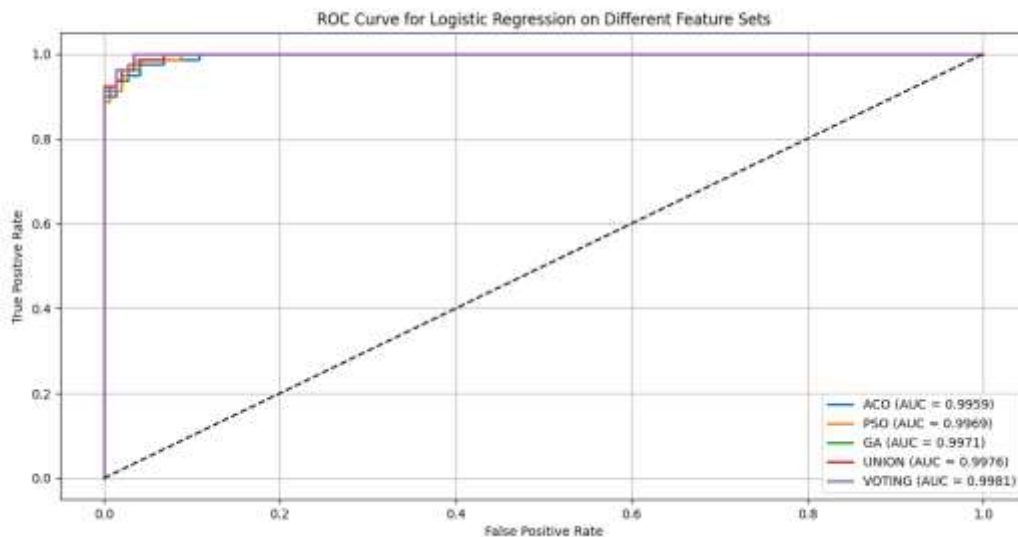


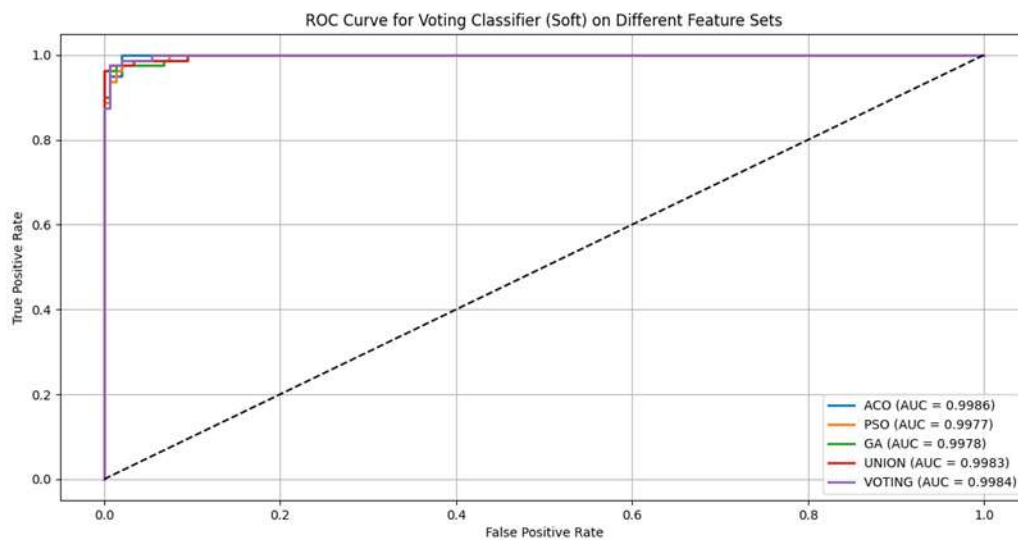
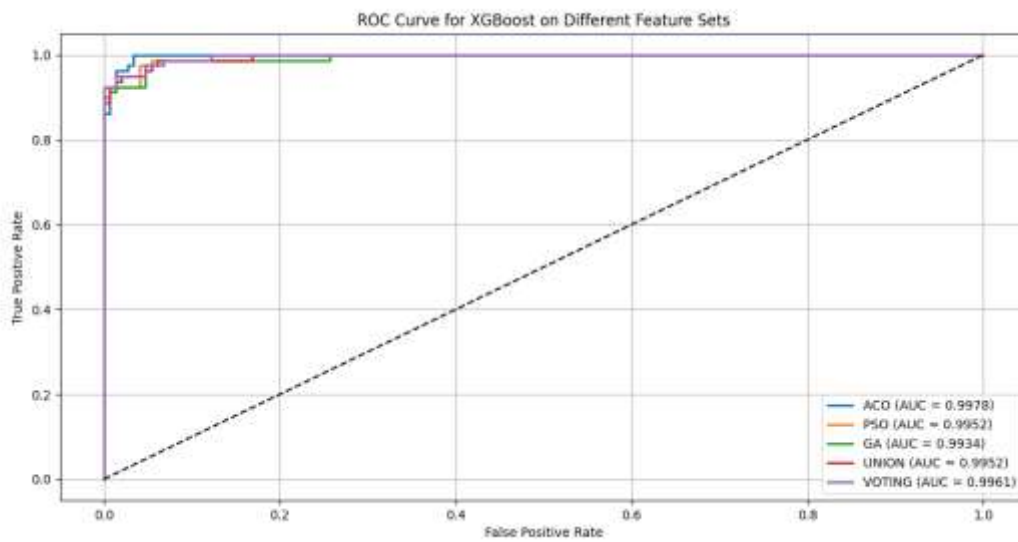
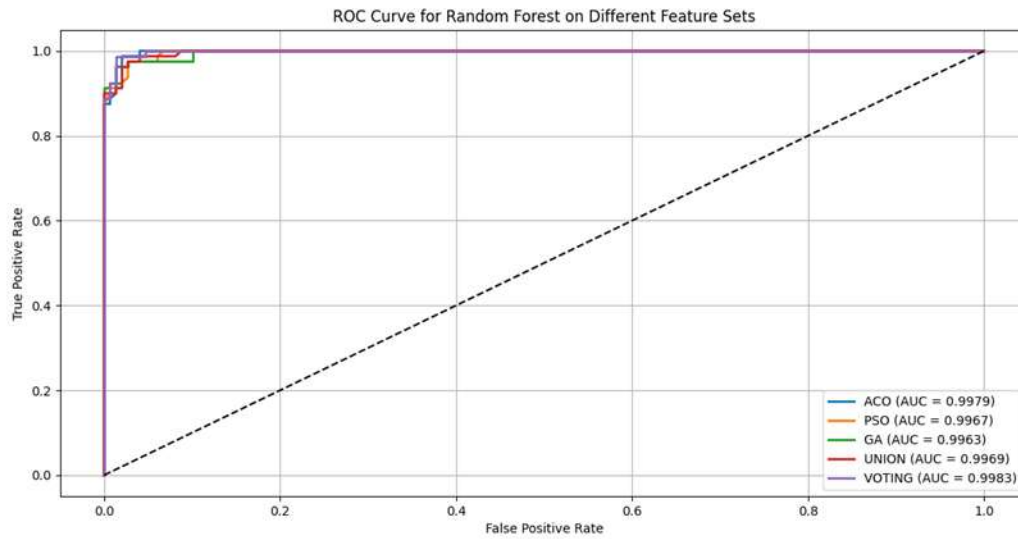
Fig. 2.1 Comparison of F1-scores among different classifiers and feature selection techniques

4.4 ROC-AUC Analysis

To further evaluate classifier performance, ROC curves were plotted for each classifier across all feature selection techniques. **Figure 3** illustrates that all models exhibit strong discriminative capabilities, with **ROC-AUC values exceeding 0.99** in most cases.

Notably, the **Voting Classifier with Voting-based features** achieved an ROC-AUC of **0.9984**, the highest among all combinations.

Fig. 3 ROC Curves of all classifiers for different feature selection techniques.

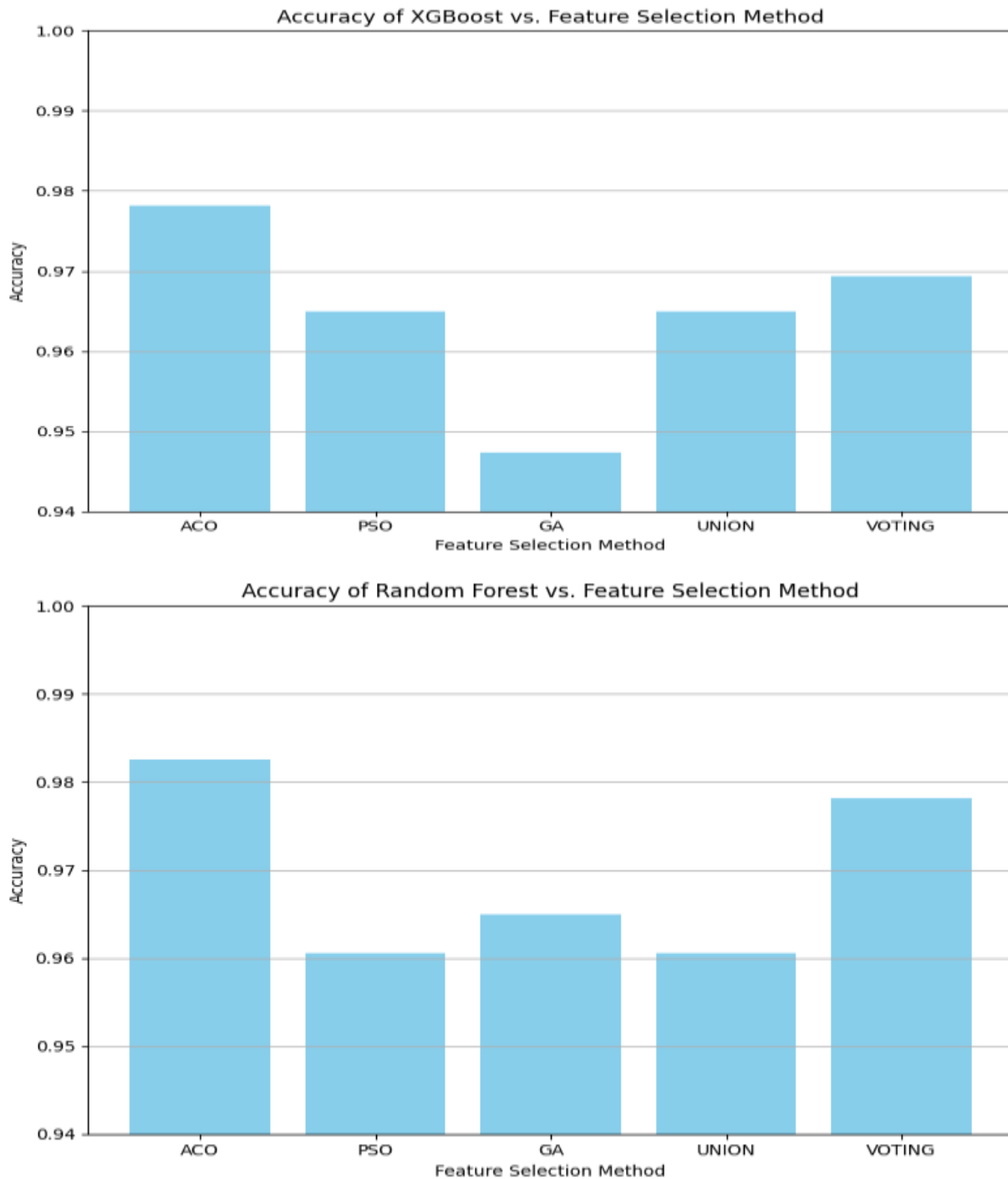


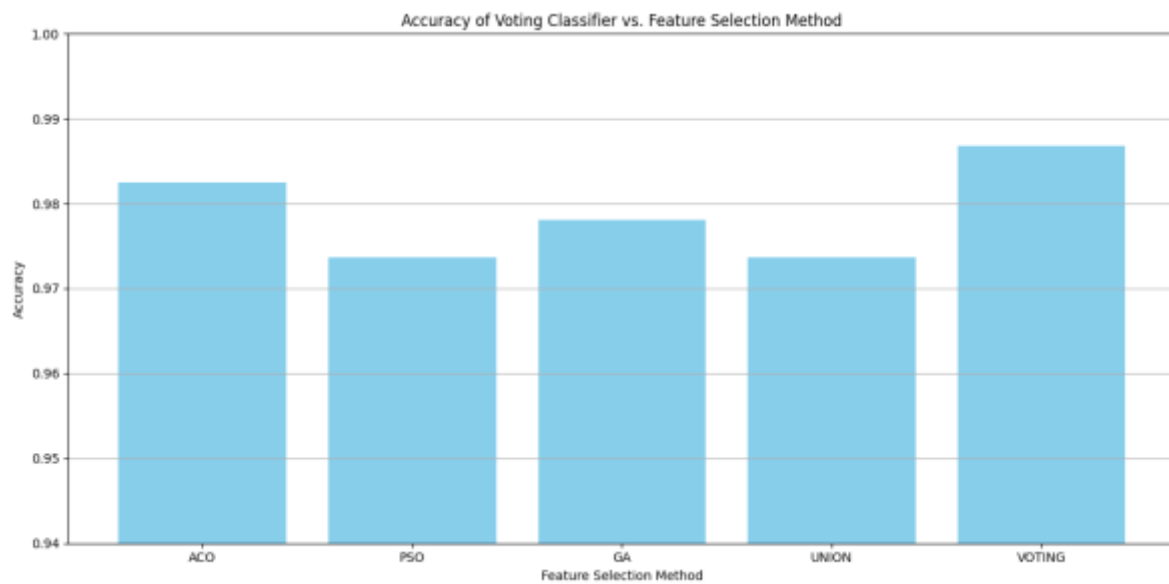
4.5 Classifier-Wise Performance Insights

Bar charts displaying performance of individual classifiers across feature selection methods provide insights into model behavior:

- **XGBoost** performed best with ACO-selected features.
- **Random Forest** showed consistent performance across all methods.
- **Voting Classifier** maintained high scores across all feature sets, demonstrating the strength of ensemble learning.

Fig. 4 Classifier-wise accuracy comparison with different feature selection techniques.

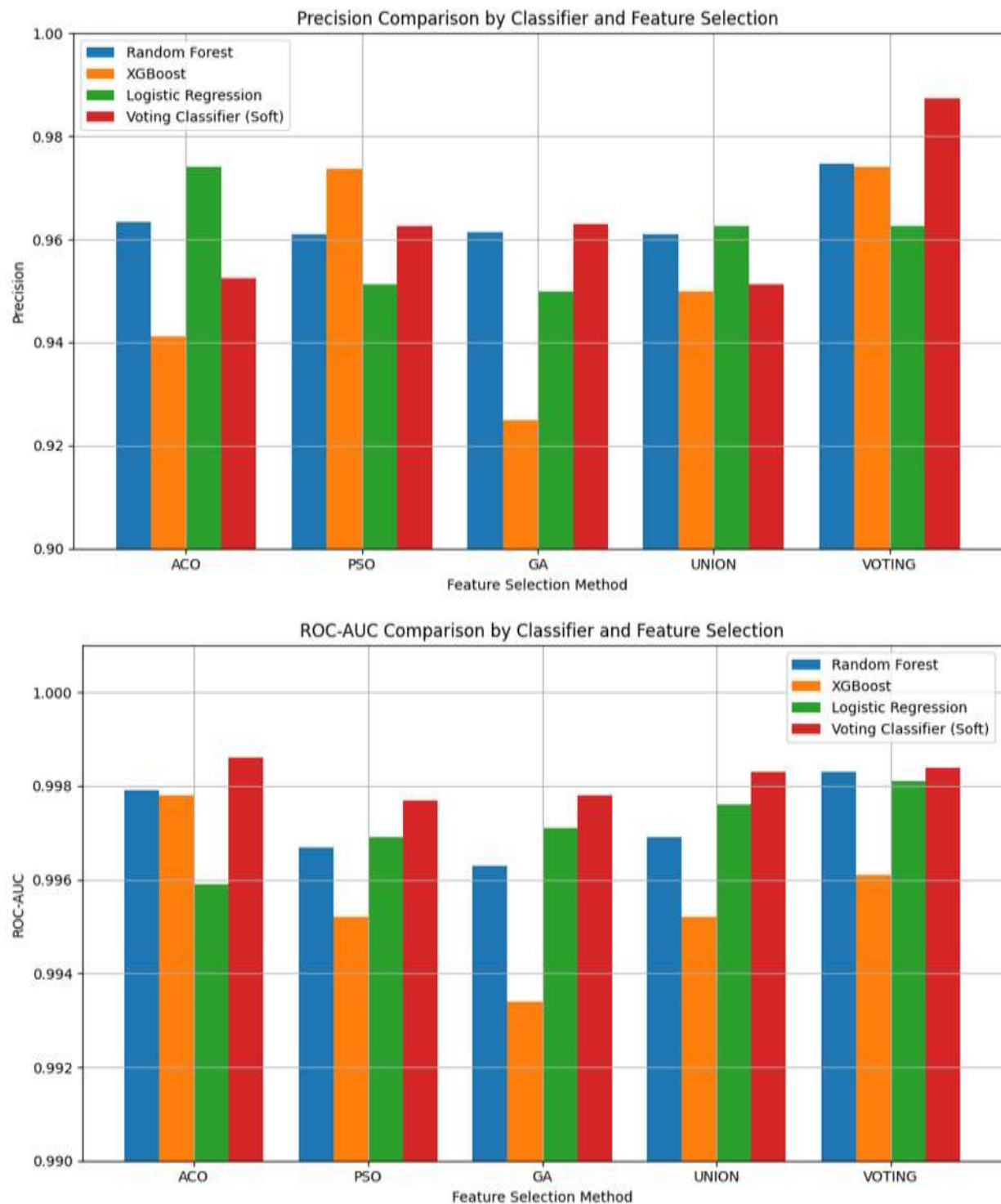




4.6 Performance Comparison Summary

Table 1 summarizes the performance outcomes—Accuracy, Precision, Recall, F1-Score, and ROC-AUC—of different classifiers applied with various feature selection strategies.

Figure 5: Precision and ROC-AUC comparison.



Key Observations:

- The **Voting Classifier with Voting-based features** consistently outperformed others with highest accuracy (98.68%), F1-score, and ROC-AUC.
- **ACO** and **PSO** feature sets also yielded strong results.
- **Union-based features**, while slightly lower in accuracy, still provided competitive performance.
- Among individual classifiers, **XGBoost** and **Random Forest** outperformed Logistic Regression in most cases.

Table 1 Comprehensive performance comparison.

Feature Selection Method	Classifier	Accuracy	Precision	Recall	F1-Score	Roc-Auc
Aco	Random Forest	0.9825	0.9634	0.9875	0.9753	0.9979
	Accuracy Xgboost	0.9781	0.9412	1	0.9697	0.9978
	Logistic Regreation	0.9693	0.974	0.9375	0.9554	0.9959
	Voting Classifier	0.9825	0.9524	1	0.9756	0.9986
Pso	Random Forest	0.9605	0.961	0.925	0.9427	0.9967
	Accuracy Xgboost	0.9649	0.9737	0.925	0.9487	0.9952
	Logistic Regreation	0.9737	0.9512	0.975	0.963	0.9969
	Voting Classifier	0.9737	0.9625	0.9625	0.9625	0.9977
Ga	Random Forest	0.9649	0.9615	0.9375	0.9494	0.9963
	Accuracy Xgboost	0.9474	0.925	0.925	0.925	0.9934
	Logistic Regreation	0.9649	0.95	0.95	0.95	0.9971
	Voting Classifier	0.9781	0.963	0.975	0.9689	0.9978
Union(Aco+Pso+Ga)	Random Forest	0.9605	0.961	0.925	0.9427	0.9969
	Accuracy Xgboost	0.9649	0.95	0.95	0.95	0.9952
	Logistic Regreation	0.9737	0.9625	0.9625	0.9625	0.9976
	Voting Classifier	0.9737	0.9512	0.975	0.963	0.9983
Voting(Aco+Po+Ga)(Hard)	Random Forest	0.9781	0.9747	0.9625	0.9686	0.9983
	Accuracy Xgboost	0.9693	0.974	0.9375	0.9554	0.9961
	Logistic Regreation	0.9737	0.9625	0.9625	0.9625	0.9981
	Voting Classifier (Soft)	0.9868	0.9873	0.975	0.9811	0.9984

Comparison of Classifiers using Different Feature Selection Techniques.

5 Discussion

The experimental results highlight several key takeaways:

- **Ensemble feature selection**, especially **majority voting**, enhances model performance by retaining robust, high-quality features.
- **Combining classifiers** in a **Voting Classifier** improves model robustness and reduces bias from individual algorithms.
- The **proposed hybrid ensemble model** outperforms single-optimizer or single-classifier systems across all key metrics.

This approach offers a strong foundation for building reliable medical diagnosis systems capable of supporting early and accurate breast cancer detection.

6 Conclusion

A hybrid ensemble framework is proposed in this study, combining optimization-based feature selection with soft voting for effective breast cancer classification. Three widely recognized metaheuristic algorithms—Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), and Genetic Algorithm (GA)—were applied to the Wisconsin Breast Cancer Dataset (WBCD) to perform dimensionality reduction by identifying the most significant features. To make the selected features more stable and meaningful, we combined the outputs of the feature selection methods using two approaches: Union and Majority Voting.

We applied several classifiers including Random Forest, XGBoost, Logistic Regression, and a soft voting ensemble to evaluate classification performance. The experimental analysis showed that the proposed framework reached a peak accuracy of 98.68%, while also delivering consistently high values for precision, recall, F1-score, and ROC-AUC. The voting-based feature selection and ensemble classification consistently outperformed individual methods, showcasing the benefits of combining multiple models and optimizers.

7 References

1. Chaurasia V, Pal S (2017) A novel approach for breast cancer detection using data mining techniques. *Int J Comput Appl* 161(8):1–4. <https://doi.org/10.5120/ijca2017913211>
2. Bharati S, Podder P, Mondal MRH (2020) Artificial neural network based breast cancer screening: a comprehensive review. *Biomed Signal Process Control* 61:102027. <https://doi.org/10.1016/j.bspc.2020.102027>
3. Karthik S, Srinivasa Perumal R, Chandra Mouli PVSSR (2018) Breast cancer classification using deep neural networks. In: Anuncia SM, Wiil UK (eds) *Knowledge computing and its applications*. Springer, Singapore, pp 227–241. https://doi.org/10.1007/978-981-10-6680-1_17
4. Han L, Yin Z (2022) A hybrid breast cancer classification algorithm based on meta-learning and artificial neural networks. *Front Oncol* 12:1042964. <https://doi.org/10.3389/fonc.2022.1042964>
5. Al-Saedi AH, Albukhnefis AL, Al-Shammary D, Al-Asfoor M (2022) Extended Particle Swarm Optimization for feature selection of high-dimensional biomedical data. *Concurrency Comput Pract Exper* 34(10):e6776. <https://doi.org/10.1002/cpe.6776>
6. Fallahzadeh O, Dehghani-Bidgoli Z, Assarian M (2018) Raman spectral feature selection using ant colony optimization for breast cancer diagnosis. *Lasers Med Sci* 33:1799–1806. <https://doi.org/10.1007/s10103-018-2544-3>
7. Karimi K, Ghodratnama A, Tavakkoli-Moghaddam R (2023) Two new feature selection methods based on learn-heuristic techniques for breast cancer prediction: a comprehensive analysis. *Ann Oper Res* 328:665–700. <https://doi.org/10.1007/s10479-022-04933-8>
8. Yang F, Xu Z, Wang H, Sun L, Zhai M, Zhang J (2024) A hybrid feature selection algorithm combining information gain and grouping particle swarm optimization for cancer diagnosis. *PLoS ONE* 19(1):e0290332. <https://doi.org/10.1371/journal.pone.0290332>
9. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI (2015) Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J* 13:8–17. <https://doi.org/10.1016/j.csbj.2014.11.005>
10. Deng X, Li M, Deng S, Wang L (2022) Hybrid gene selection approach using XGBoost and multi-objective genetic algorithm for cancer classification. *Med Biol Eng Comput* 60:663–681. <https://doi.org/10.1007/s11517-021-02476-x>
11. Sowan B, Eshtay M, Dahal K, Qattous H, Zhang L (2023) Hybrid PSO feature selection-based association classification approach for breast cancer detection. *Neural Comput Appl* 35:5291–5317. <https://doi.org/10.1007/s00521-022-07950-7>
12. Sahoo M, Bag A, Mallick SP (2024) Effect of subset on classification accuracy of breast cancer detection. *Journal of Harbin Engineering*.
13. Ameta D (2017) Ensemble classifier approach in breast cancer detection and malignancy grading: a review. *Int J Manage Policy Inf Commun Technol (IJMPICIT)* 8(1):17–26. <https://doi.org/10.5121/ijmpict.2017.8102>