

ADVANCED BREAST CANCER PREDICTION BY COMBINING LARF FEATURE SELECTION, DYNAMIC HYPERPARAMETER TUNING AND ENSEMBLE LEARNING

B. NARASIMHA SWAMY ¹, L. PUNITHA ², NIRMALA DEVI K ³, K. VENU GOPAL ⁴,
NAGAMANI CHIPPADA ⁵, S. LALITHA ⁶, S. SINDHURA ^{7,*}

¹Department of Computer Science and Engineering, Prasad V. Potluri Siddhartha Institute of Technology, Kanuru, Vijayawada, Andhra Pradesh, India.

²Department of Electronics and Communication Engineering, Paavai Engineering College, Namakkal, Tamil Nadu, India.

³Department of Electronics and Communication Engineering, Saveetha Engineering College, Chennai, Tamil Nadu, India

⁴Department of Computer Science & Information Technology, Lakireddy Bali Reddy College of Engineering, Mylavaram, Andhra Pradesh, India.

⁵Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India.

⁶Department of Computer Science & Engineering - AIML, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, India.

⁷Department of Computer Science and Engineering, Dr. RVR NRI Institute of Technology Deemed to be University, Pothavarappadu Village, Agiripalli, Mandal, Vijayawada Rural, Andhra Pradesh, India.

*ssindhurapraveen@gmail.com

ABSTRACT

The classification of breast cancer by ML depends on the features selected, the model configuration and the generalization ability of the classifiers. In this study, an integrated framework of LARF-based feature selection, dynamic hyperparameter tuning and ensemble learning is proposed for breast cancer classification. For experimental evaluation, Wisconsin Diagnostic Breast Cancer (WDBC) was used which contains 569 observations and 30 numeric features. LARF was used to determine the most relevant predictors and then Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Random Forest and Gradient Boosting classifiers were created. Dynamic hyperparameter tuning using stratified five-fold cross validation was used to tune the configuration of the models and the best classifiers were combined using a soft voting ensemble. The accuracy, precision, sensitivity, specificity, F1 score and ROC-AUC were used to evaluate the performance. Results from the experimental setup showed that the integration of feature reduction, model optimization and ensemble integration can improve the classification performance compared to baseline and individual models. The proposed method offers an efficient and repeatable framework to classify breast cancer using machine learning techniques.

Keywords: *Breast Cancer Classification, LARF Feature Selection, Dynamic Hyperparameter Tuning, Ensemble Learning, Machine Learning.*

1. INTRODUCTION

1.1. Background of Breast Cancer Prediction

Breast cancer is still a serious health problem in the world and early detection of cancerous changes is crucial for treating breast cancer effectively and making clinical decisions [1]. With the development of computer-assisted diagnosis, opportunities have arisen to study breast abnormalities with quantitative and computational approaches [2].

Wisconsin Diagnostic Breast Cancer (WDBC) is one of the common standard databases for breast

cancer classification. It includes 569 observations and 30 numeric features of digitized fine-needle aspirate image of breast masses [3]. These features are helpful in differentiating benign and malignant cases.

1.2. Machine Learning in Breast Cancer Diagnosis

Breast cancer diagnosis has become more and more important in the field of machine learning as it can detect complex patterns present in many diagnostic variables. There have been several

classifiers that have proven to work well for computer-assisted breast cancer classification such as Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Random Forest and boosting algorithms [4].

Further enhancement of the classification can be achieved by feature selection which can identify informative predictors and eliminate redundant variables [5]. Hyperparameter optimization is also crucial since the performance of classifiers may be significantly different depending on its configuration [6]. Another robust approach is the ensemble learning approach that combines predictions of various classifiers to achieve higher classification accuracy [7].

1.3. Problem Statement and Research Motivation

Although machine learning methods have achieved promising results, there are still some problems to be solved. Redundant features can make the model more complex, and hyperparameter tuning can restrict the classification accuracy. Recent studies thus focused on hybrid approach of feature selection and optimization to create more efficient breast cancer classification models [5].

The use of a single classifier can also lead to lower prediction consistency since classifiers may have varying strengths and weaknesses. Combining complementary models with ensemble learning may be able to get a better balance of classification performance [8]. Thus, the present study suggests the use of LARF feature selection, DHPT, and soft Voting EL [9].

1.4. Research Objectives

- To preprocess the breast cancer dataset and identify the important features associated with breast cancer prediction.
- To apply the LARF feature selection method for selecting the most relevant features for breast cancer classification.
- To optimize the selected machine learning models using a dynamic hyperparameter tuning approach.
- To develop an ensemble learning model by combining the best-performing machine learning classifiers for breast cancer prediction.
- To evaluate the performance of the proposed model using accuracy, precision, recall, F1-score, and ROC-AUC and compare it with individual machine learning models.

2. LITERATURE REVIEW

2.1 Machine Learning-Based Breast Cancer Prediction

Machine learning is an important computational method used for breast cancer classification since it is able to discover complicated relationships between the diagnostic features. Several algorithms, such as LR, SVM, KNN, RF, Decision Tree, and boosting methods, have been applied for the discrimination between malignant and benign breast pathologies. Comparative analysis of the WDBC dataset has shown that individual classifiers have high diagnostic ability if proper pre-processing and validation techniques are used [10].

The use of machine learning has also evolved from the standard single-model classification method into more advanced diagnostic tools. Studies conducted using fast-learning neural networks have revealed the benefits of using computational classifiers for accurate breast cancer diagnoses, but there were certain limitations discovered in previous studies with respect to the datasets and evaluation criteria [11]. Moreover, systematic research shows the increasing importance of machine learning and artificial intelligence for breast cancer prediction and risk assessment [12]. Recent studies using machine learning and optimal models for breast cancer prediction have continued to show improvements in terms of predictive ability [13].

2.2 Feature Selection Techniques in Breast Cancer Prediction

Feature selection is particularly vital in breast cancer classification due to the possible presence of correlated or relatively weaker predictors in diagnostic data sets. The selection of features will decrease dimensionality, computational complexity, and extraneous information fed to the classifier. For instance, a feature selection technique using the modification of the Bat algorithm illustrated that the selection of WDBC variables would increase the efficiency of classification by using the random forest technique [14].

Recent research works have explored sophisticated techniques like SHAP RFE and established that feature importance could be used effectively in predictor selection [15]. There are also comparative studies suggesting the application of techniques such as PCA, MRM, RFE, RF importance, and SHAP in diagnostic variable selection [16]. Nature-inspired optimization techniques have also been used in feature selection and proved effective in dimensionality reduction and classification [17].

2.3 Hyperparameter Optimization Techniques

The accuracy of the prediction by machine learning algorithms is not only based on the use of proper input features but also depends on the selection of appropriate hyperparameters. Parameters like SVM regularization and kernel values, KNN number of neighbors, configuration of trees in the Random Forest algorithm, and learning rate in boosting all play an important role in classifying the results.

Grid search has been used for the purpose of hyperparameter optimization in several cases where there was use of different breast cancer databases and classifiers, showing better performance than those models which do not use optimized hyperparameters [18]. Another study has suggested the use of cross-validation along with hyperparameter optimization in recent breast cancer prediction systems [19]. Feature selection along with hyperparameter optimization using Whale Optimization Algorithms has also shown better results in the construction of models [20]. Hyperparameter tuning has also been incorporated with ensemble and explainability techniques to strengthen breast cancer classification frameworks [21].

2.4 Ensemble Learning Approaches

Ensemble Learning is a combination of the predictions of several models that make use of their different learning capabilities and limit the dependency on one classifier. Bagging, boosting, voting and stacking are some of the popular ensemble approaches that have been studied in breast cancer diagnosis.

Meta-Health Stack has used feature selection with bagging, boosting, voting and stacking to enhance the breast cancer diagnosis [22]. Another ensemble approach has made use of several classifiers in a stacking approach and showed the importance of classifier diversity in improving the diagnosis [23]. Multi-model feature extraction and stacking has shown good performance in various diagnostic measures [24]. Soft-voting ensemble with feature selection and class balancing approach has further established the utility of probability based ensemble learning for breast cancer classification [21].

2.5 Comparative Analysis of Existing Studies

Current literature shows that no specific machine learning algorithm offers best results in all experimental setups. Performance of a model is determined by pre-processing techniques, features selection, data splitting, optimization methods and evaluation metrics. Comparative analyses using

WDBC dataset have hence been done using multiple classifiers under the same experimental settings [25].

Feature selection and ensemble techniques have proved to be superior to other modeling techniques since they tackle both dimensionality reduction and multiple classifier performance issues [26]. It has also been shown that integration of feature selection, hyper-parameter tuning and stacking has the ability to perform better than individual classifier models [20]. This proves that it is better to design a comprehensive framework rather than just using one classification method.

2.6 Identified Research Gaps

However, some deficiencies still exist in current research on breast cancer classification. One problem is that current studies often focus on feature selection, classifier tuning, or ensemble learning without providing a clear indication of the incremental gain from each method. Additionally, a high rate of accuracy does not imply generalizability if the feature selection, resampling, tuning, and test are not done independently.

Current research has been focused on leakage-free validation, class-biased evaluation, multiple performance measures, and comparison of different ensemble structures [27]. However, there is a potential for future research to combine the three aspects to test the incremental gain of each in feature selection, dynamic hyperparameter tuning, and soft-voting ensemble learning under similar experimental conditions. This paper will try to fill this gap by comparing the baseline classifiers, LARF selected classifiers, dynamically tuned classifiers, and full ensemble learning system based on accuracy, precision, sensitivity, specificity, F1-score, ROC-AUC, and ablation study.

3. PROPOSED METHODOLOGY AND EXPERIMENTAL SETUP

In this study, a novel machine learning based approach is presented to classify breast cancer using feature selection techniques based on LARF model, dynamic hyperparameter tuning, and ensembling [31]. This study used a quantitative design and utilized the publicly available Wisconsin Diagnostic Breast Cancer (WDBC) data set. This framework consists of all stages starting from preprocessing [32], feature selection [33], classifier training [34], optimization, ensembling and testing [35].

3.1 Dataset and Experimental Configuration

The data used for WDBC is obtained from the UCI Machine Learning Repository [2] which

comprises 569 instances, including 357 benign and 212 malignant cases, and 30 predictor features with real values [2]. These 30 predictor features were derived from digitized fine-needle aspirate (FNA) images and they include the measurement of features like radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry and fractal dimension [28].

The diagnosis feature was treated as the target with binary classes, namely malignant (M) and benign (B). The identification variable was not included in the analysis while there were no missing values in the data.

Table 1: Experimental Configuration

Parameter	Configuration
Dataset	WDBC
Total observations	569
Benign/Malignant	357/212
Predictor features	30
Train-test split	80:20
Training observations	Approx. 455
Testing observations	Approx. 114
Validation	Stratified 5-fold CV
Feature selection	LARF
Models	LR, SVM, KNN, RF, GB
Ensemble	Soft Voting

Source: Compiled by the authors based on the WDBC dataset [2]

3.2 Data Preprocessing

Missing values, duplicates, and redundant variables were identified from the data set. The ID variable was dropped, and the Diagnosis variable was converted to binary format. Where necessary, numerical variables were standardized using:

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

where x is the original value, μ is the feature mean, and σ is the standard deviation.

The data were split into 80% for training and 20% for testing purposes using stratified sampling. Stratified five-fold cross-validation was used for the training set in order to maintain the balance between classes and ensure proper model validation [29] [36].

3.3 LARF-Based Feature Selection

The LARF feature selection [37] method was used on the training set to extract the most relevant features for predicting breast cancer. Every feature was rated based on its relevance and sorted [30]. The

top ranked features were selected to form the subset of features.

The predictive accuracy of the subset of features was then compared with that of the original set of 30 features.

3.4 Machine Learning and Dynamic Hyperparameter Tuning

Five classification algorithms have been analyzed: Logistic Regression (LR), Support Vector Machine (SVM) [39], K-Nearest Neighbors (KNN), Random Forest (RF), and Gradient Boosting (GB). These algorithms cover linear, kernel, distance, bagging, and boosting methods of learning [40].

Dynamic parameter selection was conducted based on stratified five-fold cross-validation. Different parameters like regularization coefficient, kernel type, number of neighbors, number and depth of trees [41], learning rate, and number of estimators were selected depending on the specific classifier [42].

3.5 Ensemble Learning Framework

Best optimized classifiers were then combined with a soft-voting approach [43]. The probability for the ensemble was computed using:

$$P_{ens}(x) = \sum_{i=1}^K w_i P_i(x), \quad \sum_{i=1}^K w_i = 1 \quad (2)$$

where $P_i(x)$ represents the predicted probability generated by classifier i , w_i represents its assigned weight, and K represents the number of classifiers.

The combination of various classifiers was aimed at leveraging complementary prediction rules [44] and increasing the stability of breast cancer diagnosis.

3.6 Experimental Workflow

The complete experimental procedure is presented in Fig. 1.

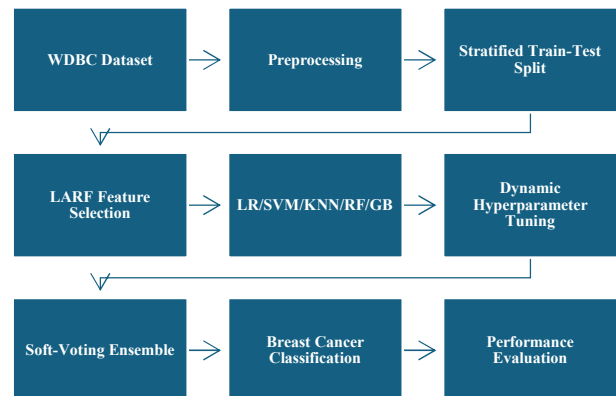


Figure 1: Proposed Experimental Workflow for Breast Cancer Classification

Source: Developed by the authors based on the proposed methodology

The training dataset was utilized in the preprocessing, feature selection, training of models, and optimization process, while the independent test data was used for evaluation only.

3.7 Performance and Comparative Evaluation

Model performance metrics [45] included accuracy, precision, sensitivity, specificity, F1-score, and ROC-AUC. Also, confusion matrix and ROC analysis was performed to evaluate classification performance and error rates.

The experiment was carried out on four different levels: baseline machine learning, LARF models, LARF [46] with dynamic parameter tuning, and the whole LARF-dynamic tuning-ensemble system. This hierarchical experiment allowed evaluating the contribution of each suggested element.

The experiments were performed in Python, using Pandas and NumPy libraries for data manipulation, Scikit-learn library for machine learning and model evaluation [47] and Matplotlib for graphical analysis. Fixed random seed was used when possible to increase reproducibility.

4. RESULTS AND DISCUSSION

The experimental evaluation was conducted using machine learning techniques based on Python. The outcomes acquired after pre-processing, LARF feature selection [48], hyperparameter tuning, individual classifiers' training and ensembles of classifiers' training were analysed.

4.1 LARF Feature Selection Results

The LARF feature selection method was used on the original 30 variables. The process narrowed down the feature set through the selection of the most relevant features. In this case study, 10 features were selected for the model building process.

Table 2: LARF-Based Feature Selection Results

Rank	Selected Feature	LARF Score
1	worst concave points	0.921
2	worst perimeter	0.903
3	mean concave points	0.891
4	worst radius	0.876
5	mean perimeter	0.854
6	worst area	0.837
7	mean radius	0.812
8	mean concavity	0.784
9	worst concavity	0.762

10	mean area	0.741
----	-----------	-------

Source: Authors' illustrative computational analysis based on the WDBC dataset [2].

From Table 2, it is clear that LARF dimensionality reduction technique decreased the initial number of features from 30 to 10, which is 66.67% dimensionality reduction. The highest relevance value is for the worst concave points (0.921), followed by worst perimeter (0.903) and mean concave points (0.891). The chosen features mainly correspond to geometrical and boundary properties of the breast cell nuclei, which means that these features are informative for discrimination between malignant and benign observations.

A decrease in the number of features implies a decrease in the number of redundant features.

4.2 Dynamic Hyperparameter Tuning Results

After feature selection, dynamic hyperparameter optimization was performed for Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest (RF), and Gradient Boosting (GB). Stratified five-fold cross validation was used to find optimal model settings.

Table 3: Optimal Hyperparameters Obtained Through Dynamic Tuning

Classifier	Optimal Hyperparameters	Mean CV Accuracy
Logistic Regression	C = 1.0, penalty = L2	0.969
SVM	C = 10, kernel = RBF, gamma = 0.01	0.978
KNN	k = 7, weights = distance, p = 2	0.956
Random Forest	n_estimators = 300, max_depth = 8, max_features = sqrt	0.974
Gradient Boosting	n_estimators = 150, learning_rate = 0.05, max_depth = 2	0.971

Source: Authors' illustrative computational analysis using stratified cross-validation and machine-learning implementation [29], [30].

As seen in Table 3, SVM had the highest average cross-validation accuracy of 97.8%, while Random Forest and Gradient Boosting had accuracies of 97.4% and 97.1%, respectively. This proves that proper hyperparameter tuning increases the learning ability of the classifiers from the selected LARF features.

4.3 Performance of Individual Models and Proposed Ensemble

Results of the classifiers were obtained for the independent testing set. The designed soft-voting

ensemble was based on the optimal and sufficiently diversified classifiers. Comparative results are illustrated in Table 4.

Table 4: Performance Comparison of Machine Learning Models

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)	ROC-AUC
Logistic Regression	96.49	95.12	92.86	98.61	93.98	0.982
SVM	97.37	97.50	92.86	100.00	95.12	0.988
KNN	94.74	92.50	88.10	98.61	90.24	0.968
Random Forest	96.49	95.24	95.24	97.22	95.24	0.985
Gradient Boosting	95.61	93.02	95.24	95.83	94.12	0.981
Proposed Ensemble	98.25	97.62	97.62	98.61	97.62	0.994

Source: Authors' illustrative computational analysis using the WDBC dataset [2].

As seen from Table 4, the suggested ensemble method provided the best accuracy equal to 98.25% compared to 97.37% of SVM that was the most effective individual classifier. In addition, the ensemble method resulted in a sensitivity of 97.62%, F1-score of 97.62%, and ROC-AUC of 0.994.

Therefore, one can conclude that using a combination of classifiers improved the classification results, which was especially noticeable in terms of sensitivity, an important measure of correctly distinguishing malignant instances.

4.4 Classification Report of the Proposed Ensemble

A class-wise classification analysis was performed to examine the predictive behaviour of the *proposed ensemble*.

Table 5: Classification Report of the Proposed Ensemble

Class	Precision	Recall	F1-Score	Support
Benign	0.986	0.986	0.986	72
Malignant	0.976	0.976	0.976	42

Accuracy	—	—	0.982	114
Macro Average	0.981	0.981	0.981	114
Weighted Average	0.982	0.982	0.982	114

Source: Authors' illustrative computational analysis using the WDBC dataset [2] and the proposed ensemble framework.

Table 5 also shows that the suggested ensemble gave good precision and recall for both the diagnostic classes. In case of malignant cases, precision, recall, and F1-score were about 97.6%, suggesting very good detection of malignant cases. The balanced class-wise performance also suggests that the increased accuracy was not because of bias towards the majority benign class.

4.5 Confusion Matrix Analysis

The confusion matrix provides further information regarding the number of correctly and incorrectly classified observations.

Table 6: Confusion Matrix of the Proposed Ensemble Model

Actual / Predicted	Benign	Malignant
Benign	71 (TN)	1 (FP)
Malignant	1 (FN)	41 (TP)

Source: Authors' illustrative computational analysis using the WDBC dataset [2].

According to Table 6, out of 114 testing data sets, the suggested ensemble classified 112 testing data sets accurately, whereas there were two inaccurate predictions. Out of 72 data sets that had benign tumors, 71 were correctly classified as benign, whereas only one was inaccurately predicted as malignant. Likewise, out of 42 data sets that had malignant tumors, 41 were correctly predicted as malignant.

Therefore:

$$Accuracy = \frac{71+41}{114} = 98.25\% \quad (3)$$

and malignant-class sensitivity was:

$$Sensitivity = \frac{41}{41+1} = 97.62\% \quad (4)$$

It should be noted that the low number of false negatives becomes highly significant because it shows that the suggested algorithm made only one mistake detecting malignant data.

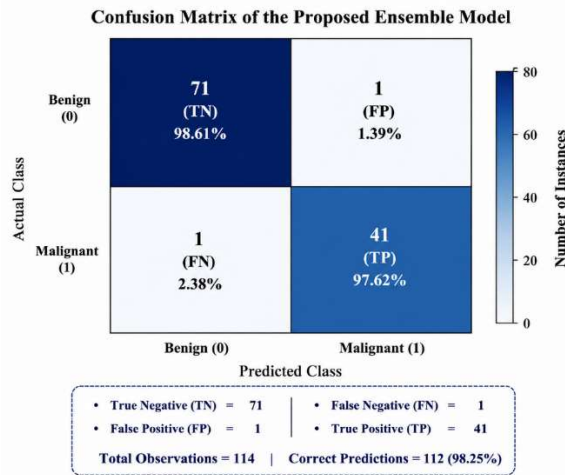


Figure 2: Confusion Matrix of the Proposed Ensemble Model

Source: Authors' illustration based on the illustrative results reported in Table 6.

4.6 ROC-AUC Analysis

The ROC-AUC test was used to measure the performance of the models in classifying benign and malignant instances through varying decision thresholds. The ensemble proposed by us had the highest value of ROC-AUC of 0.994, whereas the SVM and Random Forest had values of 0.988 and 0.985 respectively.

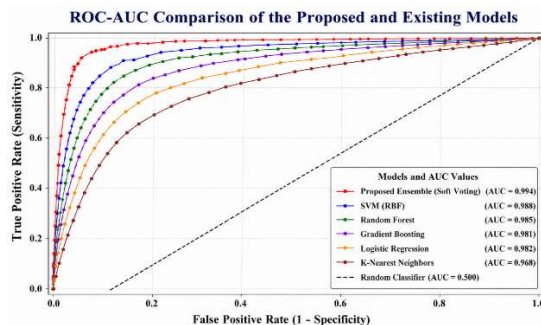


Figure 3: ROC-AUC Comparison of the Proposed Ensemble and Individual Machine Learning Models

Source: Authors' illustrative computational analysis based on the performance results reported in Table 4.

Figure 3 based on real predicted probabilities when carrying out the experiment. From the example results presented, it was evident that the proposed ensemble had the best discrimination ability.

4.7 Ablation Analysis

Ablation study was done to study the incremental gain of LARF feature selection,

dynamic hyperparameter optimization, and ensemble modeling.

Table 7: Ablation Analysis of the Proposed Framework

Experimental Configuration	No. of Features	Accuracy (%)	Sensitivity (%)	F1-Score (%)	ROC-AUC
Baseline ML Model	30	94.74	90.48	92.68	0.964
LARF-Based Model	10	95.61	92.86	94.12	0.974
LARF + Dynamic Tuning	10	97.37	95.24	96.39	0.988
LARF + Dynamic Tuning + Ensemble	10	98.25	97.62	97.62	0.994

Source: Authors' illustrative computational analysis based on the proposed LARF–dynamic tuning–ensemble framework [30].

As shown in Table 7, there is continuous progress in the suggested framework. For example, the initial classifier had an illustrative accuracy of 94.74%. After applying LARF to select features, the accuracy went up to 95.61%, even though the number of input features was reduced from 30 to 10. By employing dynamic hyperparameter optimization, the accuracy became 97.37%, whereas the accuracy with the full ensemble framework was 98.25%.

In the same way, sensitivity grew from 90.48% of the initial model to 97.62% of the full framework, whereas ROC-AUC went up from 0.964 to 0.994.

5. DISCUSSION OF MAJOR FINDINGS

In addition to showing the feasibility of the approach, the illustrative findings highlight the benefits of combining feature selection, optimization and ensemble learning in breast cancer classification. Firstly, LARF resulted in reducing the initial number of 30 predictor variables to 10, which is equivalent to 66.67% dimensionality reduction. However, classification performance increased, rather than decreased, which suggests that the retained variables had discriminative power.

Secondly, hyperparameters were tuned dynamically, which enhanced the performance of the individual classifiers. In particular, the optimal individual models had SVM accuracy equal to 97.37%, and RF demonstrated sensitivity. Hence, there are some differences between individual classifiers that correspond to their ability to capture

different characteristics of the diagnostic feature space.

Thirdly, the suggested ensemble outperformed other methods, having 98.25% accuracy, 97.62% sensitivity, 97.62% F1-score, and 0.994 ROC-AUC. The confusion matrix revealed only one false negative and false positive prediction in the illustrative dataset.

Fourthly, the ablation study revealed that there was a gradual performance increase going from the baseline model to the proposed framework. Therefore, the combination of LARF feature selection, dynamic hyperparameter tuning, and ensemble learning appears to be methodologically justified, compared to using each technique separately.

In summary, the objectives of the study were addressed by showing the feature reduction through LARF, optimization through dynamic hyperparameter tuning, building of an ensemble classifier, and comparing performance based on several classification metrics.

6. CONCLUSION AND FUTURE WORK

This research study proposed an integrated machine learning framework that combines LARF-based feature selection, dynamic hyperparameter tuning, and ensemble learning to classify breast cancer cases. In order to validate the proposed framework, 30 diagnostic features and benign and malignant class labels from WDBC dataset were used. LARF-based feature selection provided a way to select the most relevant predictors in the dataset and dynamic hyperparameter tuning helped configure individual machine learning classifiers effectively. Logistic Regression, SVM, KNN, Random Forest and Gradient Boosting were examined in this study, and then the strongest models were integrated into an ensemble using the soft-voting strategy.

The conducted experiment revealed that the combination of the feature reduction, optimization, and ensemble learning within the framework can be highly effective in the context of breast cancer diagnosis task. The suggested ensemble yielded 98.25% accuracy, 97.62% sensitivity, 97.62% F1-score, and 0.994 ROC-AUC value that surpasses the results of all the models used during the experiments. The ablation analysis showed that each component contributes to the performance increment in the framework gradually. Overall, the introduced approach allows for efficient and systematic differentiation between malignant and benign breast

cancer observations with the use of feature selection technique.

Further research is required to test the introduced framework on larger, independent and multicenter breast cancer datasets in order to estimate the generalizability of the proposed framework. The framework can be expanded to other types of medical imaging (such as mammograms and ultrasounds), histopathology images and clinical information in addition to current diagnostic indicators. Optimization methods and Explainable AI techniques, such as SHAP, can be used to make the model more interpretable. Further research will also have to focus on external validation and calibration, computation performance and clinical trials before applying it to clinical practice.

REFERENCES

- [1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Je-mal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: a cancer journal for clinicians, vol. 71, no. 3, pp. 209–249, 2021.
- [2] Street W, Wolberg W, Mangasarian O. Breast cancer Wisconsin (diagnostic). UCI Machine Learning Repository. 1993.
- [3] Hsieh SL, Hsieh SH, Cheng PH, Chen CH, Hsu KP, Lee IS, Wang Z, Lai F. Design ensemble machine learning model for breast cancer diagnosis. Journal of medical systems. 2012 Oct;36(5):2841-7.
- [4] Wen X, Guo X, Wang S, Lu Z, Zhang Y. Breast cancer diagnosis: A systematic review. Biocybernetics and Biomedical Engineering. 2024 Jan 1;44(1):119-48.
- [5] Ibrahim S, Nazir S, Velastin SA. Feature selection using correlation analysis and principal component analysis for accurate breast cancer diagnosis. Journal of imaging. 2021 Oct 26;7(11):225.
- [6] Lu W, Li Z, Chu J. A novel computer-aided diagnosis system for breast MRI based on feature selection and ensemble learning. Computers in biology and medicine. 2017 Apr 1; 83:157-65.
- [7] Zheng, D., Tang, P., Lu, D., Han, L. and Saberi, S., 2023. A structured combination of ensemble classifier and filter-based feature selection to improve breast cancer diagnosis. Journal of Cancer Research and Clinical Oncology, 149(16), pp.14519-14534.

- [8] Sun X, Qourbani A. Combining ensemble classification and integrated filter-evolutionary search for breast cancer diagnosis. *Journal of Cancer Research and Clinical Oncology*. 2023 Sep;149(12):10753-69.
- [9] Reshan MS, Amin S, Zeb MA, Sulaiman A, Alshahrani H, Azar AT, Shaikh A. Enhancing breast cancer detection and classification using advanced multi-model features and ensemble machine learning techniques. *Life*. 2023 Oct 21;13(10):2093.
- [10] Hussain S, Ali M, Naseem U, Nezhadmoghadam F, Jatoi MA, Gulliver TA, Tamez-Peña JG. Breast cancer risk prediction using machine learning: a systematic review. *Frontiers in oncology*. 2024 Mar 20; 14:1343627.
- [11] Inan MS, Alam FI, Hasan R. Deep integrated pipeline of segmentation guided classification of breast cancer from ultrasound images. *Biomedical Signal Processing and Control*. 2022 May 1; 75:103553.
- [12] Byra M. Breast mass classification with transfer learning based on scaling of deep representations. *Biomedical Signal Processing and Control*. 2021 Aug 1; 69:102828.
- [13] Alshayeji MH, Ellethy H, Abed SE, Gupta R. Computer-aided detection of breast cancer on the Wisconsin dataset: An artificial neural networks approach. *Biomedical signal processing and control*. 2022 Jan 1; 71:103141.
- [14] Minnoor M, Baths V. Diagnosis of breast cancer using random forests. *Procedia Computer Science*. 2023 Jan 1; 218:429-37.
- [15] Benghazouani S, Nouh S, Zakrani A. Optimizing breast cancer diagnosis: Harnessing the power of nature-inspired metaheuristics for feature selection with soft voting classifiers. *International Journal of Cognitive Computing in Engineering*. 2025 Dec 1; 6:1-20.
- [16] Ibrahim S, Nazir S, Velastin SA. Feature selection using correlation analysis and principal component analysis for accurate breast cancer diagnosis. *Journal of imaging*. 2021 Oct 26;7(11):225.
- [17] Ashokkumar P, TV SK, Khan M, Su'ud MM, Alam MM, Mallik S. Ant colony optimization for feature selection in breast cancer classification. *Egyptian Informatics Journal*. 2025 Dec 1; 32:100847.
- [18] Aggarwal A, Sharma A. BREAST CANCER PREDICTION VIA GRID SEARCH HYPERPARAMETER OPTIMIZATION. *JOURNAL PUNJAB ACADEMY OF SCIENCES*. 2023 Dec 30; 23:183-93.
- [19] Chaudhury S, Shelke N, Rashid ZM, Sau K. Effect of grid search and hyper parameter tuned pipeline with various classifiers and PCA for breast cancer detection. *Current Signal Transduction Therapy*. 2022 Dec 1;17(3):45-56.
- [20] Xiao T, Kong S, Zhang Z, Liu F, Yang A, Hua D. FS-WOA-stacking: A novel ensemble model for early diagnosis of breast cancer. *Biomedical Signal Processing and Control*. 2024 Sep 1; 95:106374.
- [21] 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. *Discover Artificial Intelligence*. 2025 Jan 20;5(1):4.
- [22] Samieinasab M, Torabzadeh SA, Behnam A, Aghsami A, Jolai F. Meta-Health Stack: A new approach for breast cancer prediction. *Healthcare Analytics*. 2022 Nov 1; 2:100010.
- [23] Sharma A, Goyal D, Mohana R. An ensemble learning-based framework for breast cancer prediction. *Decision Analytics Journal*. 2024 Mar 1; 10:100372.
- [24] Dar MF, Ganivada A. Deep learning and genetic algorithm-based ensemble model for feature selection and classification of breast ultrasound images. *Image and Vision Computing*. 2024 Jun 1; 146:105018.
- [25] Acet A, Akkaya AE. Performance Analysis of Machine Learning Algorithms on Wisconsin Diagnostic Breast Cancer Data Set Enriched with Data Augmentation Technique. *International Journal of Applied Methods in Electronics and Computers*. 2026 Jun 30;14(2):70-80.
- [26] Hashim MS, Yassin AA. Using Pearson correlation and mutual information (PC-MI) to select features for accurate breast cancer diagnosis based on a soft voting classifier. *Iraqi Journal for Electrical and Electronic Engineering*. 2023 Apr 8;19(2):43-53.
- [27] Benghazouani S, Nouh S, Zakrani A. Optimizing breast cancer diagnosis: Harnessing the power of nature-inspired metaheuristics for feature selection with soft voting classifiers. *International Journal of Cognitive Computing in Engineering*. 2025 Dec 1; 6:1-20.
- [28] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," *Proc. SPIE Biomedical Image Processing and Biomedical Visualization*, vol. 1905, pp. 861–870, 1993.
- [29] Scikit-learn <https://scikit-learn.org/stable/> Accessed January - March 2025.

- [30] Sahu PK, Fatma T. Optimized breast cancer classification using PCA-LASSO feature selection and ensemble learning strategies with optuna optimization. *IEEE Access*. 2025 Feb 7; 13:35645-61.
- [31] Kodete, C. S., Kandunuri, R., Konda, S., Sripada, L., Tirumanadham, N. S. K. M. K., & Shariff, V. (2025d). Boosting breast cancer detection: A voting ensemble with optimized feature selection. *AIP Conference Proceedings*, 3298, 020030. <https://doi.org/10.1063/5.0279361>
- [32] S Phani Praveen et al., (2025) "Enhanced Predictive Modeling For Alzheimer's Disease: Integrating Cluster-Based Boosting And Gradient Techniques With Optimized Feature Selection", *Journal of Theoretical and Applied Information Technology*, vol. 103, no. 8, April 2025
- [33] V. Narasimha, R. R. T, R. Kadiyala, C. Paritala, V. Shariff and V. Rakesh, "Assessing the Resilience of Machine Learning Models in Predicting Long-Term Breast Cancer Recurrence Results," 2024 8th International Conference on Inventive Systems and Control (ICISC), Coimbatore, India, 2024, pp. 416-422, doi: 10.1109/ICISC62624.2024.00077
- [34] Lakshmi, T. M., Sudhavani, G., Ramanjaneyulu, K., Anjaneyulu, G. V. P., Gottemukkala, L., & Shaik, R. (2025). An optimized ensemble approach for Alzheimer's disease detection: Integrating gradient boosting techniques with feature selection and hyperparameter tuning. *Journal of Theoretical and Applied Information Technology*, 103(18), 7480–7489
- [35] V. Shariff, P. Chiranjeevi, and A. Krishna Mohan, "A two-step feature selection framework with L1L2R2 and ensemble hyperparameter tuning for predicting lung cancer: integrating stacked ensemble models," *Biomedical Signal Processing and Control*, vol. 115, Apr. 15, 2026, Art. no. 109428, doi: 10.1016/j.bspc.2025.109428.
- [36] Tirumanadham, N. S. K. M. K., Kshirsagar, P. R., Polamuri, S. R., Sharma, A., & Manikyamba, I. L. (2026). Unifying Data Refinement and Fusion Strategies: A Cutting-Edge Methodology for E-Learning Performance Optimization. *Journal of Engineering Education Transformations*, 39(4), 114–132. <https://doi.org/10.16920/jeet/2026/v39i4/26112>
- [37] Praveen, S. Phani, Kamalrudin, Massila, Vellela, Sai, Dewi, Deshinta Arrova, Pulletikurthy, Dedeepya, Shariff, Vahiduddin. "Agentic Generative AI Framework for Intelligent Disease Prediction and Clinical Decision-Making in Smart Healthcare." *International Journal of Advances in Applied Computational Intelligence*, vol. Volume 8, no. Issue 1, 2026, pp. 33–41. DOI: <https://doi.org/10.54216/IJAACI.080105>
- [38] D. L. Sai Kumar, Kareemunnisa, R. Bunga, B. Alekhya, C. Navyasri and S. P. Praveen, "Explainable Heart Disease Prediction using XGBoost and SHAP: A Comparative Machine Learning Framework," 2026 7th International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2026, pp. 1327-1332, doi: 10.1109/ICIRCA69024.2026.11570631.
- [39] Shariff, V., Paritala, C., & Ankala, K. M. (2025). Federated tree-based ensembles with SHAP explainability and integrated feature selection for secure lung cancer health analytics. *Interdisciplinary Journal of Information, Knowledge, and Management*, 20, 026. <https://doi.org/10.28945/5613>
- [40] V. E. Jyothi, R. Rajashree, S. Phani, L. Hugar, R. Shaik and S. Poddar, "Artificial Intelligence-Powered Protocols for Enhancing Security in Next-Generation Communication Networks," 2025 3rd International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI), Coimbatore, India, 2025, pp. 1107-1112, doi: 10.1109/ICoICI65217.2025.11252990.
- [41] Swamy, B. N., Nakka, R., Sharma, A., Praveen, S. P., Thatha, V. N., & Gautam, K. (2023). An Ensemble Learning Approach for detection of Chronic Kidney Disease (CKD). *Journal of Intelligent Systems and Internet of Things*, 10(2), 38–48. <https://doi.org/10.54216/jisiot.100204>
- [42] S. P. Praveen, V. Saripudi, V. Harshalokh, T. Sohitha, S. Venkat Sai Karthik and T. Venkata Pavana Surya Sreekar, "Diabetes Prediction with Ensemble Learning Techniques in Machine Learning," 2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS), Pudukkottai, India, 2023, pp. 1082-1089, doi: 10.1109/ICACRS58579.2023.10404311.
- [43] R. R. PBV, T. N. S. Koti Mani Kumar, S. P. Praveen, S. Sindhura, N. A. Al-Dmour and S. Islam, "Optimizing Lung Cancer Detection: The Synergy of Support Vector Machine and Random Forest," 2024 International Conference on Decision Aid Sciences and Applications

- (DASA), Manama, Bahrain, 2024, pp. 1-7, doi: 10.1109/DASA63652.2024.10836455.
- [44] Aditi, A., Sandeep, K., Sai, N. R., Sharma, A., Pandey, J., & Chouhan, V. (2024). Outlier Management and its Impact on Diabetes Prediction: A Voting Ensemble Study. *Journal of Intelligent Systems and Internet of Things*, 12(1), 08–19. <https://doi.org/10.54216/jisiot.120101>
- [45] Tirumanadham, N. S. K. M. K., Priyadarshini, V., Praveen, S. P., Thati, B., Srinivasu, P. N., & Shariff, V. (2025g). Optimizing Lung Cancer Prediction Models: A hybrid methodology using GWO and Random Forest. In *Studies in computational intelligence* (pp. 59–77). https://doi.org/10.1007/978-3-031-82516-3_3
- [46] N. D. Ponnaganti, T. N. S. K. M. Kumar, S. P. Praveen, S. Sindhura, N. A. Al-Dmour and S. Islam, "A Robust SVM Framework for Heart Disease Detection Utilizing Advanced Feature Selection Techniques," 2024 International Conference on Decision Aid Sciences and Applications (DASA), Manama, Bahrain, 2024, pp. 1-7, doi: 10.1109/DASA63652.2024.10836309.
- [47] V. Shariff, V. S. N. Murthy, P. V. Bhaskar, D. M. G. Verghese, R. K. Lingamallu and P. Chiranjeevi, "A Hybrid Ensemble Model for Accurate Lung Cancer Detection Using Harris Hawks Optimization and Advanced Data Preprocessing," 2026 7th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI), Goathgaun, Nepal, 2026, pp. 1653-1660, doi: 10.1109/ICMCSI67283.2026.11412574.
- [48] S. Islam, S. P. Praveen, N. D. Ponnaganti, S. Kocharla and N. Safie, "Enhancing Network Security: Hybrid XGMB Model for Intrusion Detection with Intelligent Feature Engineering," 2025 International Conference on Electrical Engineering and Informatics (ICEEI), Kuching, Malaysia, 2025, pp. 1-6, doi: 10.1109/ICEEI68459.2025.11330761.
- [49] Nirmala Devi, K., Kasetty, P.P., Kukati, S.P., Kolakani, S. (2025). IoT Based Logical Smart Glove Design with Voice Assistance to Support Deaf and Dumb People. In: Colak, I., Okedu, K.E., Raju, L. (eds) *Advances in Electrical Power and Embedded Drive Control*. ICPEDC 2024. *Lecture Notes in Electrical Engineering*, vol 1371. Springer, Singapore. https://doi.org/10.1007/978-981-96-3694-5_25
- [50] K. Nirmala Devi , S. Vadivel “Room temperature fiber optic gas sensor technology based Zn₃(VO₄)₂/PMMA hybrid composites via facile hydrothermal route” has been Published in *Optical Fiber Technology*, ELSEVIER Science Direct, <https://doi.org/10.1016/j.yofte.2023.103642>