Revista Internacional de Métodos Numéricos
para Cálculo y Diseño en Ingeniería

ARTICLE

A Robust Ensemble Learning Paradigm for Breast Cancer Diagnosis Based on Multi-Classifier Performance Optimization

Ali Waqar[1,*], Muhammad Jamshed Abbass[2,*], Zohaib Mushtaq[3,4], Madiha Mariam[5], Ayasha Iram[6], Muhammad Taha[7] and Muhammad Atif Jamil[8]

[1]School of Mechanical and Electrical Engineering, Quanzhou University of Information Engineering, Quanzhou, Fujian, China
[2]Faculty of Electrical Engineering, Wrocław University of Science and Technology, Wybrzeże Wyspiańskiego 27, Wrocław, Poland
[3]Department of Electrical Engineering, Yuan Ze University, Taoyuan, Taiwan; dr.zohaib-mushtaq@saturn.edu.tw
[4]Research Center of Applied Mathematics, Khazar University, Baku, Azerbaijan
[5]Fatima Jinnah Medical University, Lahore, Pakistan
[6]Nishtar Medical University, Multan, Pakistan
[7]Department of Electronic Engineering, Islamia University of Bahawalpur, Bahawalpur, Pakistan
[8]ATECHSolutions LLC, Texas , USA; Atifjamil@gmail.com
*Corresponding Authors: Ali Waqar. Email: aliwaqar@qzuie.edu.cn; Muhammad Jamshed Abbass. Email: muhammad.abbass@pwr.edu.pl

Received: 27 February 2026; Accepted: 27 July 2026

ABSTRACT: Breast cancer is among the major causes of cancer-related mortality in women, and early diagnosis is critical to improved survival. However, traditional diagnostic methods or standalone models often lack precision and have the potential to give false-negative results. This study aims to present a hyperparameter-tuned ensemble learning model for the classification of breast benign and malignant tumors based on the Wisconsin Breast Cancer (Original) dataset. Data quality was ensured through a series of preprocessing techniques: median imputation, Zscore standardization, and the Local Outlier Factor (LOF) noise reduction algorithm. The classical Machine Learning (ML) models (SVM, Random Forest (RF), XGBoost, Bagging) were compared with the Deep Learning models (VGG16 and DenseNet121). A hybridized soft-voting ensemble of the four classical classifiers (SVM, RF, XGBoost, and Bagging) was then constructed, and its hyperparameters were tuned with the OPTUNA framework using the Tree-structured Parzen Estimator (TPE), a Bayesian optimization algorithm, over 30 trials. The results of the experiments demonstrated that the Ensemble + OPTUNA model performed with an accuracy of 99.25% and a recall of 1.00 for the malignant class compared to the individual models and recent state-of-the-art methods. This is an optimized ensemble for reliable and computationally stable performance. To enhance transparency and privacy-preserving validation in clinical settings, Explainable AI (XAI) will be incorporated into future research, while Federated Learning will be enhanced to better validate the model.

KEYWORDS: Breast cancer diagnosis**;** ensemble learning**;** OPTUNA optimization**;** machine learning**;** deep learning**;** Wisconsin breast cancer dataset**;** soft-voting

1 Introduction

Breast cancer is one of the most widespread and dangerous cancer worldwide and a major health burden in terms of morbidity and mortality [1]. Recent global statistics indicate that in 2022, more than 2.3 million women were diagnosed with breast cancer, and nearly 670,000 deaths were reported worldwide [2], which depicts its significance for public health. The disease affects women across all regions, with a higher incidence in later life. Breast cancer typically arises from the uncontrolled growth of abnormal cells in breast tissue [3], forming benign or malignant tumours [4]. The most common signs and symptoms include

a palpable breast lump, changes in breast size or shape, skin dimpling, nipple discharge, and local pain [5]. If not diagnosed early, the disease can spread to lymph nodes and distant organs, increasing the risk of mortality. Radiation therapy, chemotherapy, hormonal therapy, and targeted therapy are some of the available treatment options, depending on the stage and biological features of the tumour [6]. Therefore, early and accurate diagnosis is essential to improve survival. In this context, structured clinical datasets like the Wisconsin Breast Cancer (WBC) (Original) dataset, which includes quantitative features of fine needle aspiration (FNA) of breast cell nuclei, are excellent starting points for developing machine learning-based diagnostic systems. Diagnosis using conventional methods [7], such as biopsy and clinical examination, can be reliable but is invasive, time-consuming, and requires expert interpretation. These limitations may result in inconsistent diagnoses, particularly in early diagnosis. To overcome these challenges, machine learning (ML) techniques have been increasingly applied to structured medical data for automatic classification [8]. However, several problems remain in previous studies, including overfitting, feature-selection issues, and limited generalization across different datasets. Some feature selection techniques, such as Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE), introduce inconsistencies in the preprocessing and evaluation methods, leading to unstable results. In addition, stronger and more stable predictive models may be needed, since individual classifiers may not capture some of the more complex patterns in the data. This is mainly due to the limitations of classical classifiers, which is leading diagnostic systems to adopt new hybrid methods, such as orthogonal Rademacher–Fourier moments combined with deep neural networks. These effectively capture nonlinear and complex features in biomedical images to enable highly interpretable cancer recognition [9]. Furthermore, quaternion Meixner moments (optimized with CNNs) significantly reduce dimensionality and enhance accuracy in larger and more challenging image classification tasks [10]. To address the weaknesses of traditional diagnostic approaches and single machine learning models, this study employs a supervised learning framework enhanced by ensemble learning techniques. Machine learning methods are particularly effective with structured numerical variables, where intricate non-linear relationships can be discovered that cannot be easily captured by conventional statistical methods. Single classifiers, however, are typically susceptible to weaknesses such as noise sensitivity, overfitting, and limited reproducibility across different data distributions. To overcome these issues, the proposed solution combines multiple classifiers using a soft-voting scheme, leveraging the predictive capabilities of multiple models to achieve greater robustness. The ensemble model reduces variance and improves classification reliability through the aggregation of probabilistic outputs. This strategy provides more stable and consistent predictions and is an appropriate and efficient method to improve diagnostic performance in breast cancer classification tasks. The key contributions of this work are as follows:

- Development of a weighted soft voting ensemble model that combines Random Forest (RF), a Bagging-based classifier, an SVM classifier, and XGBoost, in which a weight is assigned to each model to emphasise stronger learners, leading to enhanced accuracy, stability, and robustness in breast cancer classification.

The remainder of this paper is organized as follows. Section 2 describes related work on breast cancer classification. Section 3 presents the proposed methodology, including preprocessing, model development, and the evaluation strategy. Section 4 reports the experimental findings and performance of the proposed method against existing models. Finally, Section 5 concludes the paper and outlines possible future research directions.

2 Review of Literature

Existing progression in ML has considerably improved the quality and effectiveness of breast cancer diagnosis over structured clinical data. Many studies have been developed to address different classification methods to improve predictive performance and aid in early detection. The hybrid Swin-ViT and DeepLabV3+ multi-model transfer learning framework proposed by Rehman et al. [11] to analyse kidney carcinoma demonstrates the increasing role of state-of-the-art vision transformers in medical decision

support systems. Mamun et al. [12] (2025) explored the WBC and implemented a series of ML models SVM, KNN, Naïve Bayes (NB), and Neural Networks and implemented normalization and SMOTE as two preprocessing methods. The Neural Network obtained a better performance of 98.13% accuracy and high AUC-ROC. The WDBC dataset was utilized by Kadhim and Kamil [13] (2022) to compare eleven different ML classifiers and determined that they are the DT, KNN, SVM, MLP, and ensemble algorithms (Bagging and Extra Trees). The model that performed the best was the ERT model with a 97.36% accuracy and 96.77% F1-score. However, the study is limited to model comparison without a novel framework and lacks advanced ensemble optimization and external validation. Shahnaz et al. [14] (2017) used the WBC dataset and applied multiple ML and deep learning (DL) models, including SVM, KNN, RF, MLP, and CNN with feature selection techniques. The CNN had the best accuracy of 98.06, though the study has limitations such as the inability to have a single optimized framework and the higher computing demands of deep models without ensemble optimization. Mahmood et al. [15] (2020) examined DNNs and machine learning models applied to various datasets of breast cancer images (e.g., DDSM, MIAS, INbreast) to detect, segment, and classify breast cancer tissues, with high accuracy (97–99) in CNN models. Nissar et al. [16] (2024) used the CMMD dataset and was validated on MIAS and CBIS-DDSM, introducing a MobileNet-V3-based dual-channel attention model (MOB-CBAM). The model achieved up to 99% accuracy (98% fine-grained) but is limited by complex architecture and reliance on computationally intensive imaging data. Ibeni et al. [17] (2019) used UCI breast cancer datasets (WBC, WDBC, Breast Tissue) to compare Bayesian classifiers (NB, BN, TAN) with KNN, SVM, and DT. The BN model gets the highest accuracy of 97.28%, but study is limited to comparison only and reports challenges in scalability and high-dimensional data handling. Hossin et al. [18] (2023) used the WDBC dataset to compare eight ML models (LR, RF, KNN, DT, SVM, AdaBoost, GB, GNB) with feature selection (UFS, RFE). Logistic Regression achieved the highest accuracy of 99.12%. Mushtaq et al. [19] (2020) used the WBC and WDBC datasets and proposed an optimized KNN model with multiple distance functions and feature selection methods (L1-based and, **Chi-2**.). The model achieved up to 99.42% accuracy. Abdel-Zaher and Eldeib [20] (2016) used the WBC Dataset and proposed a hybrid deep learning model combining Deep Belief Network (DBN) with backpropagation NN (DBN-NN) for CAD-based classification. The model achieved 99.68% accuracy but is limited by high computational complexity and requires significant training time and resources. Nissar et al. [21] (2025) used the CMMD mammogram dataset and proposed a radiomics-based ML framework with metaheuristic optimization (DE & GWO) and LGBM classifier for breast cancer subtype detection. The model had a high accuracy of about ~81–83%. For handling multi-modal images in terms of data privacy and interpretability, Ahmed et al. [22] suggested a federated ensemble of a hierarchical Swin Transformer family and a Random Forest meta-learner. In a similar way, Munshi et al. [23] proposed an optimized ensemble approach coupled with XAI methods, achieving enhanced transparency while maintaining high diagnostic capabilities. These works encompass the ensemble optimization along with the interpretability and privacy-preserving mechanisms. Despite achieving high accuracy, many existing studies suffer from limitations such as lack of generalization, and absence of robust ensemble strategies.

3 Methodology

In this section, the overall computational model that is intended to be used in the robust classification of breast cancer is outlined. The methodology includes preprocessing of data, outlier elimination, running traditional and deep learning baselines, and constructing the proposed hyperparameter-optimal ensemble framework. The processes described in this research document are represented in the computational workflow in **Figure 1** as a block diagram of the whole experiment. This process includes the steps of data preprocessing, baseline model evaluation, and development of the OPTUNA-optimised ensemble classifier.

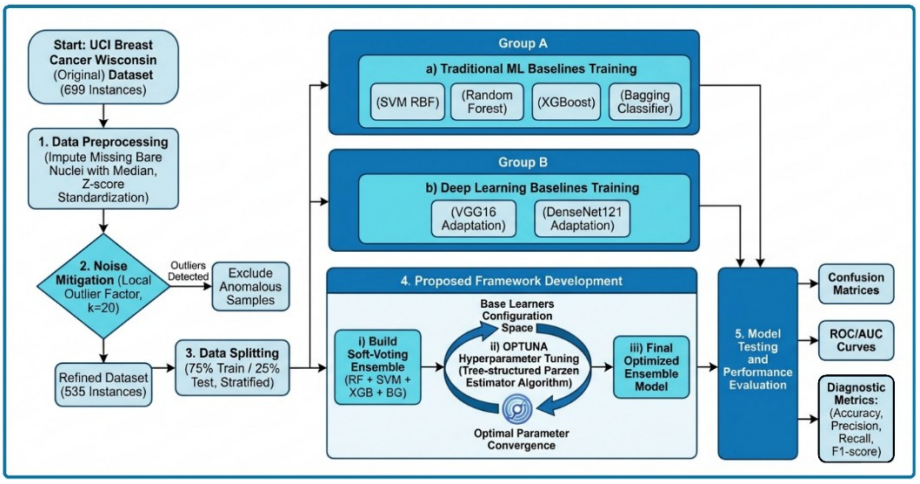


Figure 1: General Block diagram of the proposed experimental framework, illustrating the data preprocessing pipeline, baseline model evaluation, and the development of the OPTUNA-optimized ensemble classifier.

3.1 Overview of the Data Set

The data set used in the research is the Wisconsin Breast Cancer (Original) Dataset [24] that is accessed on the UCI Machine Learning Repository. It is a popular data set used in breast cancer diagnosis research and has 699 instances containing 9 predictive variables and a single target. The data consists of cytological features of breast tissue samples, and each characteristic denotes a particular attribute of the cell nuclei, given as a result of FNA of breast masses. The input characteristics are the well-spaced clumps of cells, the uniformity of size and shape, marginal adhesion, single cell size of epithelial cells, bare nuclei, bland chromatin, normal nucleoli, and mitoses. All are integer valued, with feature scores being 1–10 in increasing levels of abnormality. The target variable is discrete (approximately 2 or 4) and the values denote a tumor as benign (class = 2) or malignant (class = 4). The dataset is moderately imbalanced, with 458 benign (~65.5%) and 241 malignant (~34.5%) instances (≈1.9:1). This was addressed through stratified train–test splitting, with results reported using class-wise metrics rather than accuracy alone. One feature, Bare Nuclei, has missing values, which had been pre-processed. This dataset is of particular interest to machine learning applications due to its structured form, middle-size, and well-defined features, and it is regarded as a benchmark dataset to categorize models into cancer detection of breast cancer.

3.2 Data Preprocessing and Outlier Handling

Accordingly, a rigorous preprocessing pipeline was established to ensure the most suitable convergence of the model, and to minimize the impact of outliers. Target labels were initially mappable to standard binary nomenclature and the y labels were assigned 0 and 1 for the benign and malignant classes, respectively. For the Bare Nuclei feature only, missing data were substituted with the median from the distribution of the feature, to ensure data integrity but without inducing statistical bias. Imputation is done followed by applying the LOF algorithm to detect outlier data and eliminate them. The LOF computes the local density deviation of a data point but with respect to its k-nearest neighbors and rejects those examples that have density values that are much smaller than what is expected of their close neighbours. With the $k = 20$ chosen neighbourhood size, the unusualness of the cytological profiles that can affect the shape of the decision regions had been successfully removed. Finally, Standardization of features over the

feature space was performed using the Z score method to reduce the domination of features having bigger numeric interval on the objective functions. For a specific map x, the standardized vector x' was determined as follows:

$$x' = \frac{x - \mu}{\sigma} \tag{1}$$

where the μ is mean of sample, and the standard deviation σ. These adopted preprocessing techniques have greatly helped in enhancing the quality of the data and maintaining the stability of classification. Median imputation was chosen because it is less sensitive to outliers and is efficient in maintaining and preserving the statistical distribution of clinical features with missing data. Using LOF for outlier removal removed abnormal and noisy instances that may cause the boundaries of the decisions to be misrepresented and have a negative influence on the generalization of the model. In addition, Z-score normalization standardized feature magnitudes, preventing features with larger numerical ranges from dominating the learning process and improving convergence of distance-based classifiers such as SVM. Refined dataset was further divided into training (75%) and testing (25%) subsets by means of stratified sampling in order to maintain the distribution of classes in the original dataset.

3.3 Traditional Machine Learning Baselines

Four classical supervised learning algorithms were applied to develop a strict comparative standard of evaluation. The SVM is the first model that classifies input features into a high dimensional space to find the best separating hyperplane that maximizes the margin between classes. Through the application of Radial Basis Function (RBF) kernel, SVM is able to manage non-linear decision boundaries found in a complex clinical data. The SVM is based on the minimization of the weights and classification errors, i.e.,

$$\min_{\omega,b,\xi} \frac{1}{2} \| \mathrm{w} \|^2 + C \sum_{i=1}^{N} \xi_i \;\; subject\ to\ y_i(\omega^T \phi(x_i) + b) \geq 1 - \xi_i \tag{2}$$

In this case, regularization parameter is C, which determines the trade-off between maximization of margin and the classification error, and ξ_i are the slack variables, which are necessary to guarantee generalizability.

After the SVM, a RF classifier was used to suit intricate data interactions. This algorithm is a bagging-based ensemble algorithm that builds an enormous number of unpruned decision trees in the training stage. Each tree is trained on a different, bootstrapped subset of the original dataset, which adds required randomness and much less variance than single decision trees. In the case of classification, the final prediction $\hat{y}$ is calculated by a majority vote mechanism over all constituent trees, and is mathematically defined as,

$$\hat{y}_{RF} = mode\{h_1(x), h_2(x), \ldots, h_t(x)\} \tag{3}$$

where $h_t(x)$ represents the discrete prediction of the *t*-th independent tree.

Moreover, eXtreme Gradient Boosting (XGBoost) model has been added to take advantage of its scalability and computation efficiency (sequential tree-building). The XGBoost does not operate in parallel as RF does, but instead forms successive trees, which aim to reduce the residual errors of the prior trees, using a gradient descent algorithm to minimize a particular loss function. The regularized objective function to be minimized at step t is,

$$L^{(t)} = \sum_{i=1}^{N} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{4}$$

where l acts as a differentiable convex loss function measuring the difference between the target and the prediction, and $\Omega(f_t)$ serves as a penalty term for tree complexity to strictly prevent overfitting.

Finally, a standard Bagging Classifier (BG) was implemented to complete the baseline traditional array. It is a technique that methodically constructs a set of homogenous base estimators, usually standard unpruned decision tree, on randomly chosen subsets of the training data, with replacement. The Bagging classifier actually smoothes decision boundaries on each individual model by independently training them and combining the results of the individual distribution with others. This operation remarkably decreases the aggregate model variance and boosts predictive strength of the presence of underlying noise that can be seen in cytological feature collections.

3.4 Deep Learning Baselines

To explore such a capability of neural architectures to capture complex and hierarchical cytological features, two advanced deep convolutional neural networks VGG16 and DenseNet121 were modified to this classification task. Classical architectures are routinely used to process 2D image grids of spatial data, so first the 1D structured clinical features were made into conventional multidimensional tensor features, which resembled pseudo-images. The VGG16 network is a very deep network with 3×3 layers of convolutional filters each. A hierarchical cascade allows the network to extract increasingly hierarchical feature representations, which is extremely powerful to capture intuitively latent correlational structure across the cytological variables. In contrast, DenseNet121 model suggests a much more sophisticated and efficient connection model. The dense block is used as direct feature map as input to all the previous layers combined. This is naturally a very high connectivity which in addition makes it possible to make a rich reuse of features across the network and also appreciably decreases the vanishing gradient problem during backpropagation.

3.5 Proposed Optimized Ensemble Framework

Although the standalone models are effective, they are vulnerable to localized overfitting and variance, in the case of heterogeneous clinical data. A hybridized soft voting ensemble system is used to solve this problem, which integrates the specific predicative power of RF and SVM with XGBoost and Bagging Classifier. The ensemble of classifiers brings complementary strengths to the classification of breast cancer. Random Forest and SVM are competitive algorithms in terms of performance, with Random Forest offering noise and overfitting resistance by support vector machine due to its bagging resampling, and SVM showing the better capability to delineate the non-linear decision boundaries in high-dimensional clinical data. XGBoost has the ability to improve the prediction performance sequentially by minimizing the classification error, while Bagging smooth out prediction and boost the generalization. By integrating these heterogeneous learners in a soft voting ensemble, probabilistic decision fusion realized so that the disadvantages of the single classifiers strengthened by other classifiers, thereby enhancing the robustness, stability and diagnosis accuracy of the ensemble. The soft voting mechanism exploits the predicted probabilities of classes, unlike hard voting, which only sums the discrete labels of classes predicted by the base learners. Let $M = \{M_{RF}, M_{SVM}, M_{XGB}, M_{BG}\}$ represent the set of base learners. For an input instance x, each model M_k outputs a posterior probability $P_k(y = c \mid x)$ for class $c \in \{0, 1\}$. The ensemble calculates a weighted probability sum, defined as:

$$P_{ensemble}(y = c \mid x) = \frac{\sum_{k=1}^{4} w_k P_k(y = c \mid x)}{\sum_{k=1}^{4} w_k} \tag{5}$$

where w_k denotes the specific weight assigned to the k-th classifier. The final predicted class $\hat{y}$ is deduced by maximizing the ensemble probability:

$$\hat{y} = \arg \max_{c \in \{0,1\}} P_{ensemble}(y = c \mid x) \tag{6}$$

To allow the diagnostic potential of this ensemble to shine in full glory, the hyperparameter optimization structure OPTUNA was incorporated. Random search or standard grid search methods are computationally exhaustive and can fail to detect the global optima. However, OPTUNA exploits a Bayesian optimization algorithm, in this case, the Tree-structured Parzen Estimator (TPE). Assume that Λ is the joint space of all four constituent models. The OPTUNA framework aims at identifying the optimal hyperparameter setting $\lambda^* \in \Lambda$ minimizing the validation loss function $\mathcal{L}$:

$$\lambda^* = \arg\min_{\lambda \in \Lambda} \mathcal{L}(\lambda, \mathcal{D}_{train}, \mathcal{D}_{val}) \tag{7}$$

The TPE algorithm represents the probability density of the hyperparameters conditioned on the loss score, stating two distributions, $l(x)$ where hyperparameter configurations achieve loss below a dynamically adjusted threshold y^*, and $g(x)$ where hyperparameter configurations achieve loss above this threshold,

$$P(x \mid y) = \begin{cases} l(x) & \text{if } y < y^* \\ g(x) & \text{if } y \geq y^* \end{cases} \tag{8}$$

OPTUNA selects the best point in search space by maximizing the Expected Improvement (EI) which, in turn, is proportional to the ratio $\frac{l(x)}{g(x)}$ in an intelligent way. It numerically optimizes key parameters of the RF and XGBoost, including the number of estimators, maximum depth, and SVM C and γ parameters, and finally, the optimal ensemble configuration that can optimize the true positive rates in breast cancer diagnostics is determined.

The OPTUNA framework improved ensemble performance by identifying optimized hyperparameter combinations that enhanced classification accuracy and reduced overfitting through 30 Tree-structured Parzen Estimator (TPE) based optimization trials. Among the tuned parameters, SVM (C) and gamma influenced decision boundary flexibility, while Random Forest and XGBoost depth and estimator settings improved model stability and variance reduction. In XGBoost, learning rate and sampling parameters further enhanced generalization across heterogeneous clinical samples. Alternative ensemble strategies such as stacking or boosting-based hybrids may also achieve competitive performance. However, stacking methods often increase computational complexity and overfitting risk for relatively small datasets. Therefore, the proposed soft-voting ensemble was selected due to its simplicity, computational efficiency, interpretability, and stable generalization performance.

3.6 Performance Evaluation

Performance evaluation is used to assess the effectiveness and generalizability of the machine learning models. The accuracy, precision, recall, F1 score, confusion matrix, and the value of the area under the receiver operating characteristics (AUC-ROC) were all used to assess the individual classifiers and the proposed ensemble model in this analysis. Such metrics can take a broad account of the model's performance and indicate General statistics of classification performance, its ability to recognize the positive samples, its sensitivity to false positive and false negative rate. The proportion of correctly predicted instances over the test set called accuracy and given by:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{9}$$

TP (True Positive), TN (True Negative), FP (False Positive) and FN (False Negative) are the number of instances that were correctly and incorrectly classified respectively. Although accuracy gives general information, it may fail in the case of an uneven dataset, where the classes are not

distributed equally. Precision is the fraction of the predictively TP relative to the total positive predictions:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{10}$$

It shows that the model is able to determine the right number of positive cases out of all the projected positive cases. Recall (or sensitivity) is the rate of true positives to total actual positive cases:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

It gives the sensitivity of the model to engage a maximum number of positive samples. F1-score is the harmonic mean of precision and recall, a balanced measure of the two:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

F1-score is specifically beneficial when there is an imbalance across classes and effects the trade-off between recall and precision.

AUC-ROC curve checks the capacity of a model to differentiate between the positive and negative classes by defining the rate of True Positive (TPR) and False Positive (FPR) against one another. TPR equals:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{13}$$

and FPR as:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{14}$$

The area under the ROC (AUC) is the measure of ability by the model to discriminate, where a 0.5 (AUC) shows random guessing of the data, and (AUC) of 1.0 shows perfect classification. Increased AUC denotes increased performance about differentiating between the two classes.

In addition, Matthews Correlation Coefficient (MCC) and Precision-Recall AUC (PR-AUC) were used for more comprehensive evaluation in medical diagnosis scenarios. MCC provides a balanced assessment of classification performance and is defined as:

$$\text{MCC} = \frac{\text{TP}\times\text{TN} - \text{FP}\times\text{FN}}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{15}$$

PR-AUC evaluates the trade-off between precision and recall and is particularly useful for assessing positive-class detection performance.

To further guarantee a clinically extensive assessment, Balanced Accuracy and Cohen's Kappa coefficient were also calculated for each model. Balanced Accuracy is the average (arithmetic mean) of the sensitivity and specificity values, which is a good performance measure that caters to possible class imbalance,

$$\text{Balanced Accuracy} = \frac{(Senstivity + Specificity)}{2} \tag{16}$$

Cohen's Kappa (κ) quantifies the level of agreement between predicted and actual class labels beyond what would be expected by chance alone, and is defined as:

$$K = \frac{(p_o - p_e)}{(1 - p_e)} \tag{17}$$

where p_o denotes the observed classification accuracy and p_e represents the expected agreement by chance based on the marginal totals. A Kappa value approaching 1.0 reflects near-perfect agreement, while values near 0 indicate performance equivalent to random classification. These supplementary metrics collectively

provide a more rigorous and clinically interpretable assessment of model performance beyond accuracy alone.

3.7 Experimental Setup

The experiments of this analysis were performed on a machine with an Intel Core i9-14-th generation CPU with a high clock speed, 1 TB of hard drive, 32 GB of RAM, and 10 GB of 3080 graphic card. The system is based on the windows 11 operating system, which enables smooth integration of both CPU and GPU in various model training procedures. Particularly, the CPU was used in training conventional ML models, and the GPU was used to speed up the training of DL models, which offers considerable computational benefits to the more resource-intensive tasks. This investigation took the form of software environment on Python, and the code was run using Keras library to construct deep learning models. The development environment was jupyter notebook which provides an interactive and flexible environment to write code, manipulate data and do visualizations. The proposed framework is computationally efficient as it is based on classical machine learning models operating on structured tabular data. SVM, Random Forest, XGBoost and Bagging classifiers are low- to medium-complexity to train and allow rapid inferences once trained. Although the soft-voting ensemble raises training cost a little because it uses multiple base learners, this cost is only incurred during the training phase, and it does not affect real-time prediction. Likewise, Hyperparameter tuning using OPTUNA introduces additional preprocessing time but is only conducted during model selection, offline. Thus, the resulting optimized ensemble achieves the same inference efficiency as single models. To minimize overfitting and potential data leakage, a stratified train-test split strategy was adopted prior to model evaluation. Preprocessing and model fitting procedures were carefully applied to prevent information leakage from the testing data during training. The proposed Ensemble+OPTUNA framework demonstrates strong computational efficiency suitable for real-time clinical deployment, achieving a mean inference time of 45.82 ± 4.80 ms per individual sample and 0.43 ms per sample across the full test set of 134 instances. The results validate that although the four base classifiers are heterogeneous, the inference time is still negligible with the proposed framework, which is suitable for being integrated into real-time Clinical Decision Support Systems (CDSS).

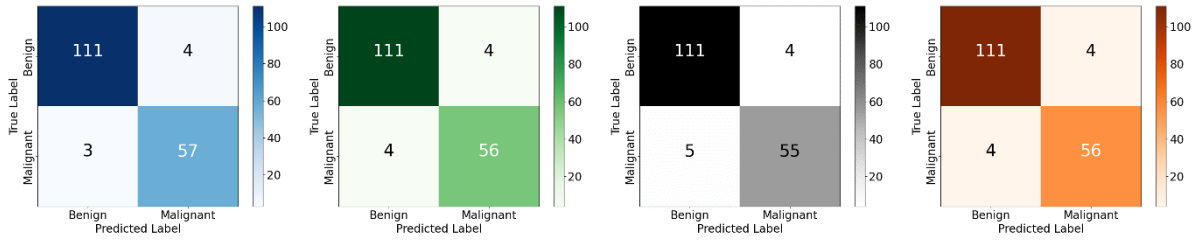
4 Results

All the outcomes of the tests performed on each of the evaluated models are presented here, with a particular focus on the capacity for breast cancer classification. This evaluation is done in three different categories namely: traditional ML baselines, DL architectures and the proposed optimized ensemble framework. There is a discussion of each category, and a visual aid using confusion matrices and Receiver Operating Characteristic (ROC) curves visualizing the situation of classification and sensitivity vs. specificity.

4.1 Performance Analysis of Traditional Machine Learning Baselines

The main experimentation step was the testing of the classical machine learning algorithms: the Support Vector Machine (SVM), the Random Forest (RF), the XGBoost (XG) and the Bagging (BG) classifiers. After the analysis of the testing set, the SVM model was the most effective stand-alone traditional classifier with all-over accuracy of 96.00%. It recorded a macro-average precision of 0.95, recall of 0.96, and F1-score of 0.96. For the benign class, the SVM achieved a precision, recall, and F1-score all standing at 0.97, while for the malignant class, it achieved a precision of 0.93, a recall of 0.95, and an F1-score of 0.94. The RF and XGBoost models yielded identical

overall test accuracies of 95.43%, alongside matching macro-average scores of 0.95 for precision, recall, and F1-score. Both models similarly achieved a precision, recall, and F1-score of 0.97 for the benign class, and 0.93 across all three metrics for the malignant class. The Bagging classifier demonstrated the lowest baseline performance, with an overall test accuracy of 94.86% and macro-average precision, recall, and F1-scores all at 0.94. For the benign class, the Bagging model recorded a precision of 0.96, a recall of 0.97, and an F1-score of 0.96, whereas its performance on the malignant class resulted in a precision of 0.93, a recall of 0.92, and an F1-score of 0.92. These baseline models differ in individual confusion matrices, which can be graphically shown in **Figure 2** specifying the distribution of their actual positive and actual negative selection and their misclassifications respectively. Moreover, the derived ROC curves as shown in **Figure 3** remark the discriminating ability of each of the baselines with the SVM having the best area under the curve as compared to the traditional models.In order to study the effectiveness of NN-based methods, the VGG16 and DenseNet121 pre-trained architectures were modified and trained and tested only on the testing set. VGG16 had a good testing accuracy of 96.57 and a test loss reading of 0.1210. Its test metrics were detailed with a precision of 0.9355, recall of 0.9667 and F1-score of 0.9508. Moreover, VGG16 showed superb separation of classes with AUC-ROC of 0.9933. Other error and correlation values of VGG16 were a Mean Squared Error (MSE), Mean Absolute Error (MAE) of 0.0343, and a robust Hamming loss of 0.0343 as well as a MCC of 0.9248 and a Cohen Kappa of 0.9245. On the other hand, the testing accuracy of DenseNet121 was 95.43% and the test loss was higher at 0.1728. The testing set resulted in a precision of 0.9483, recall of 0.9167 and F1-score of 0.9322. DenseNet121 AUC-ROC was noted to be 0.9919.

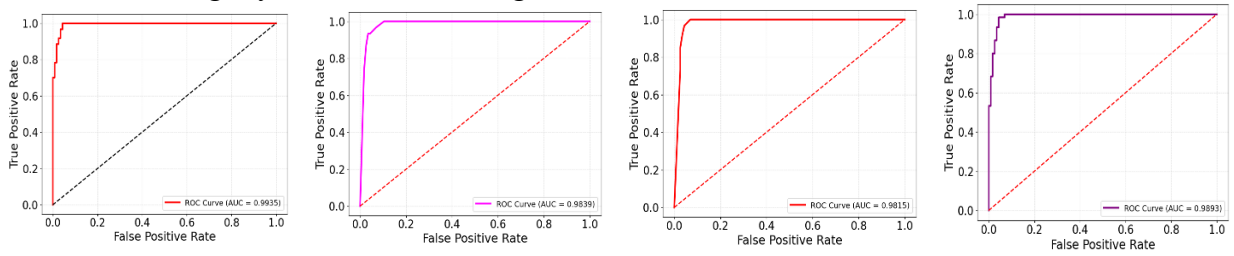


(a) (b) (c) (d)

Figure 2: Confusion matrices of the standalone traditional ML baselines on the testing set: (a) SVM model, (b) RF Model, (c) Bagging Classifier, and (d) XGBoost.

4.2 Diagnostic Evaluation of Deep Learning Architectures

All error measures of this model consisted of an MSE, MAE, and Hamming loss equal to 0.0457 with MCC of 0.8980 and Cohen Kappa equal to 0.8977. The confusion matrices shown in **4(a)(b)** and ROC curves of these deep learning baselines illustrated in **Figure 4(c)(d)** offer a graphical overview of their diagnostic characteristics, as they demonstrate the high sensitivity of VGG16 and a slightly increased false-negative rate in DenseNet121.



(a) **(b)** **(c)** **(d)**

Figure 3**:** Receiver Operating Characteristic (ROC) curves for various models. (a) SVM model, (b) RF model, (c) Bagging classifier, and (d) XGBoost model. Each curve represents the model's TPR against the FPR.

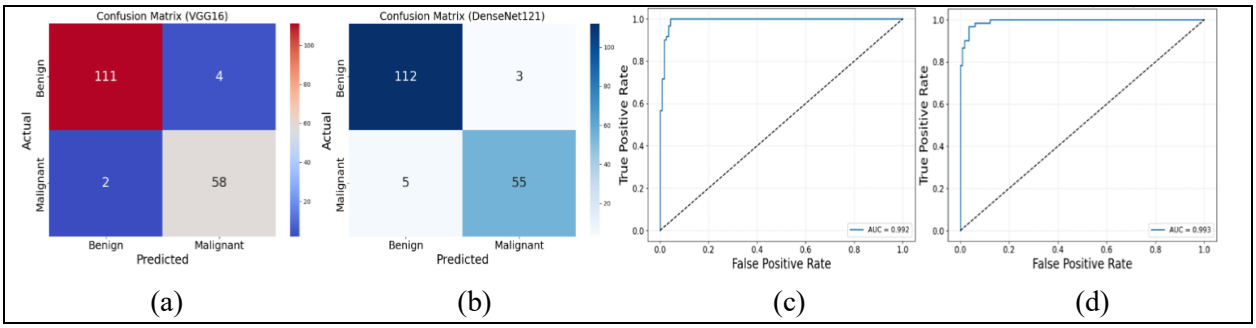


Figure 4**:** Confusion matrices for binary classification (Benign vs. Malignant): (a) VGG16 and (b) DenseNet121 and ROC curve of (c) DenseNet121 classifier, and (d) VGG16 model on the test set.

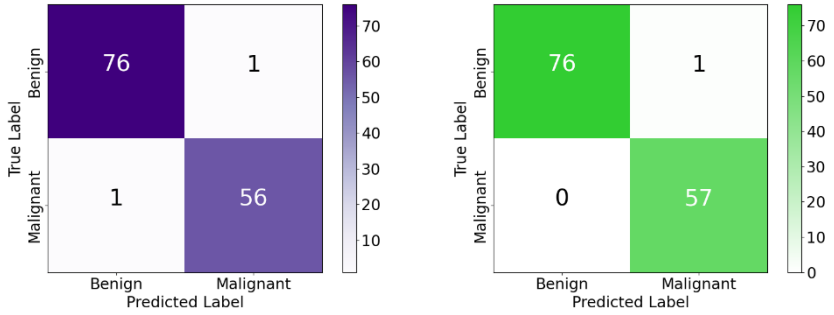
4.3 Efficacy of the Proposed Optimized Ensemble Framework

To overcome the weaknesses of individual models and gain as much diagnostic accuracy as possible, an ensemble learning model based on RF, SVM, XGBoost, and the Bagging classifier was tested on a fined testing set of 134 instances. The baseline accuracy of the standard, unoptimized ensemble model was impressive: 98.51%. This setup produced macro-average scores of 0.98 in terms of precision, recall and F1-score. In the benign subclass, it attained an almost perfect precision, recall, and F1-score of 0.99 whereas the malignant subclass recorded a score of 0.98 in the same three. Further enhance this performance, OPTUNA framework was used to tune the hyperparameters of underlying base learners systematically. With these tuned parameters, the recommended ensemble with OPTUNA model was the most effective predictor in the study with an excellent testing accuracy of 99.25%. Macro-average precision, recall and F1-score scores of the optimized framework were 0.99. More importantly, it had a perfect precision of 1.00 (benign), and perfect recall of 1.00 (malignant), and this gave a F1-score of 0.99 respectively. This structural advantage is confirmed by the confusion matrices of the standard and optimized ensemble models as depicted in **Figure 5** which shows almost all classification errors have been eliminated. The relevant ROC curves seen in **Figure 6** demonstrate a near-perfect true-positive curve, which decisively represents the conclusion that the optimized ensemble framework that has been based on the first instance of the optimal tuning, provide unsurpassed robustness and diagnostic performance when it comes to breast cancer classification. In order to give a clear and direct comparison between the diagnostic capability of each of the experimented algorithms, the testing set performances of the traditional ML baselines, DL architectures and proposed ensemble frameworks are synthesized into Table 1. The point of this summary is that the classification performance is gradually improving, and the best final results are shown by both the highest level of accuracy, precision, recall, MCC, PR-AUC, and by the highest level of the F1-score, which the ensemble model optimized by OPTUNA crosses.

Table 1: Comprehensive Performance Summary of All Evaluated Models (Testing Set).

Model	Acc. (%)	Precision (Macro)	Recall (Macro)	F1-Score (Macro)	MCC	PR-AUC	Cohen's Kappa	Balanced Acc.
Bagging Classifier	94.86	0.94	0.94	0.94	0.8855	0.9343	0.8854	0.9409
XGBoost	95.43	0.95	0.95	0.95	0.8986	0.9768	0.8986	0.9493
RF	95.43	0.95	0.95	0.95	0.8986	0.9468	0.8986	0.9493

SVM	96	0.95	0.96	0.96	0.9117	0.9868	0.9116	0.9576
DenseNet121	95.43	0.95	0.92	0.93	0.8980	--	0.8977	0.9453
VGG16	96.57	0.94	0.97	0.95	0.9248	--	0.9245	0.9659
Ensemble (unoptimized)	98.51	0.98	0.98	0.98	0.9695	0.9978	0.9645	0.9847
Proposed Study	99.25	0.99	0.99	0.99	0.9849	0.9978	0.9848	0.9935



(a) (b)

Figure 5: Confusion matrices comparing the performance of the ensemble approaches on the testing set: (a) simple soft-voting ensemble learning model, and (b) proposed ensemble learning model optimized with OPTUNA.

A class-wise error analysis of the misclassified samples further reveals the clinical superiority of the proposed framework. Among the standalone traditional classifiers, the Bagging Classifier produced the highest number of false negatives (FN = 5, FNR = 8.33%), indicating a greater risk of missed malignant diagnoses, while SVM performed best among individual models (FN = 3, FNR = 5.00%). DenseNet121 similarly exhibited a high FNR of 8.33% despite competitive overall accuracy. In contrast, the proposed Ensemble+OPTUNA model produced only one misclassification a single false positive achieving a complete elimination of false negatives (FNR = 0.00%), which is the most clinically critical outcome in breast cancer diagnosis, as any missed malignant case may critically delay life-saving treatment.

4.4 Ablation Study

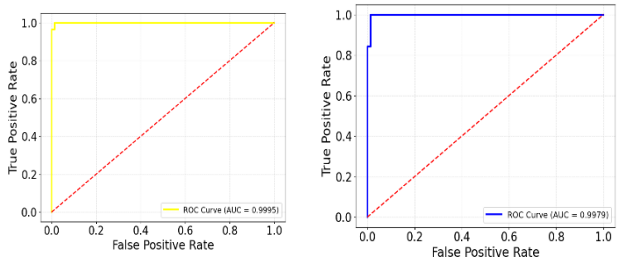
A systematic ablation study was performed by training each base model independently and comparing its performance with the performance of the proposed ensemble configuration to quantify the contribution of the individual base classifiers in the proposed ensemble framework. The results of this analysis are presented in Table 2. As illustrated in the table, none of the individual classifiers is as diagnostic as the entire ensemble. Using the Bagging Classifier alone, it has an accuracy of 94.86%, MCC = 0.8855 and FNR = 8.33%, which is a high risk of incorrectly predicting a malignant diagnosis. Random Forest model has an accuracy of 95.43% with a FNR of 6.67% and SVM model has the highest accuracy of 96.00% with a FNR of 5.00% among all the standalone models. The progressive improvement becomes evident when the four classifiers are combined into the unoptimized soft voting ensemble giving accuracy of 98.51%, the MCC of 0.9695 and the FNR of 1.75%. Bayesian hyperparameter optimization using the further incorporation of OPTUNA-based proposed model leads to a model that achieves a peak accuracy of 99.25%, an MCC of 0.9849, and a complete elimination of false negative results with an FNR of 0.00%. These results conclusively show the importance of each component of the ensembles in enhancing the overall diagnostic strength of the proposed framework, and that the use of heterogeneous classifiers with systematic hyperparameter optimization is crucial to obtaining clinically acceptable breast cancer classification performance.

Table 2: Ablation Study—Contribution of Each Ensemble Component.

Configuration	Classifiers Included	Acc.	FNR	Improvement
Bagging only	BG	94.86	8.33%	--
XGBoost only	XGB	95.43	6.67%	--
RF only	RF	95.43	6.67%	--
SVM only	SVM	96.00	5%	Baseline
Ensemble (unoptimized)	RF+SVM+XGB+BG	98.51	1.75%	+2.51%
Proposed Study	**RF+SVM+XGB+BG +OPTUNA**	**99.25**	**0.00%**	**+3.25%**

4.5 Discussion

The results of this research indicate the high diagnostic accuracy of the suggested framework of the optimized ensemble using OPTUNA to classify breast cancer. Although the standalone traditional machine learning models and VGG16 types of deep learning architectures had high accuracy baselines of between 94.86 and 96.57, they were also prone to misclassification. Common drawbacks of single classifiers include sensitivity to noise, overfitting, and less predictive accuracy on a variety of data distributions. As summarized in Table 3, the proposed ensemble framework consistently outperformed all baselines, with OPTUNA-based optimization providing further improvement. The principle clinical advancement of the optimized ensemble model is its perfect recall (1.00) of the malignant group, which means that it functionally eliminates false-negative predictions. Minimizing false negative is essential in the case of oncology; an underserved diagnosis in the form of a malignant condition may defer life-saving interventions and having a drastic effect on patient survival. Beyond recall, a comprehensive diagnostic utility assessment further validates the clinical reliability of the proposed framework. Compared to the traditional classifiers as SVM (FPR = 3.48%), the Ensemble+OPTUNA model has a low False Positive Rate (FPR = 1.30%), with only 1 out of 77 benign cases misclassified as malignant cases. The model has a Positive Predictive Value (PPV) of 0.9828, meaning that 98.28% of the time when the model predicts a tumour is malignant, it is indeed malignant, and a Negative Predictive Value (NPV) of 1.0000, which is the model's ability to accurately identify a benign tumour when it predicts it. This offers the unique advantage of having no false negatives, low false positive rate, and near-perfect PPV/NPV, making the proposed framework a very good diagnostic tool for clinical use.



(**a**) (**b**)

Figure 6: ROC curves evaluated on the testing set for the advanced frameworks: (a) simple soft-voting ensemble model, (b) proposed OPTUNA-optimized ensemble model

The model proposed has very competitive and strong performance compared to recent literature that uses the Wisconsin Breast Cancer data. For instance, Mamun et al. got 98.13% accuracy using Neural Networks and Hossin et al. [18] could achieve 99.12% accuracy using

Logistic Regression. The structured tabular ensemble enables direct inspection of per-classifier probabilities, which facilitates trust among clinicians in the AI-assisted diagnosis made by the tabular ensemble, unlike SHAP or LIME dependent post-hoc explanation pipelines for deep models.

4.6 Comparative Analysis with Existing Studies

To confirm the efficacy, robustness and clinical viability of the framework proposed, its diagnostic performance was strictly checked against the recent methodologies that used datasets (WBC, WBCD and WDBC). This comparative analysis is summarized in Table 3, which outlines the datasets used, the decision adopted approaches, and their end-of-rank classification accuracy to the literature as a whole. Other articles have experimented widely with different machine learning and deep learning designs of breast cancer classification; such as Rovshenov et al. [25] applied the Artificial neural Network (ANN) to the WBC dataset, with an exceptionally impressive accuracy score of 99.00, and similar work by Mamun et al. [12] used standard Neural Network (NN) on the same data, where the classification accuracy is 98.13%. Exploring the Wisconsin Breast Cancer Diagnostic (WBCD) dataset, Jony et al. [26] analysed the WDBC dataset with Feedforward Neural Networks (FNN) and CNN, both networks demonstrated the same accuracy of 96.49%. In another method that reduces features, Hossin et al. [18] showed excellent performance of traditional machine learning in their application of Logistic Regression (LR) with Univariate Feature Selection (UFS) and Recursive Feature Elimination (RFE) in WDBC dataset, achieving a high accuracy of 99.12%. Although these available methodologies have excellent predictive quality, they often use standalone architectures, which might be computationally expensive, prone to localised overfitting, and without systematic, dynamic hyperparameter optimization. Conversely, in our analysis, we propose a hybridized soft-voting Ensemble Learning framework optimized with the system of Bayesian hyperparameter optimization via OPTUNA. Our proposed framework achieves state-of-the-art performance relative to these existing approaches (**Table 3**), while offering a more stable and computationally efficient tabular-based diagnostic solution. The proposed Ensemble with OPTUNA framework low inference latency which is suitable for real-time structured data-based clinical decision support. OPTUNA added extra offline optimization overheads when training the model, but with little additional computation before deployment.

Table 3: Comparative performance of proposed model with Recent studies.

Authors	Dataset	Approach	Accuracy
Rovshenov et al. [25]	WBC	ANN	99
Jony et al. [26]	WDBC	FNN	96.49
		CNN	96.49
Mamun et al. [12]	WBC	Neural Network	98.13
Hossin et al. [18]	WDBC	LR with (UFS, RFE)	99.12
Our Study	**WBC**	**Ensemble Learning with OPTUNA**	**99.25**

4.7 Limitations and Future Implications

However, there are a few points to consider in the light of the high diagnostic performance obtained. The first step in the empirical analysis was to use a dataset of breast cancer cases from Wisconsin (Original), available as a single source (699 samples). Although it is a popular benchmark, its small size and single-institution basis may preclude generalisation to larger, multi-

centric and demographically diverse populations (age, ethnicity, geographic origin), with the potential to impact fairness and equitable performance across under-represented groups. The framework will be tested on additional datasets in the future, and will aim to be externally validated, using multi-institutional datasets from external institutions (e.g., METABRIC, TCGA), to assess generalization and clinical reliability. Second, the framework is purely based on formalized cytological features of the Fine Needle Aspiration (FNA) biopsy, rather than using raw imaging data, thus not a full end-to-end diagnosis system based on imaging. Future extensions will involve multimodal integration of radiological imaging, genomic biomarkers and clinical data for enhanced diagnostic completeness, integration of Explainable AI techniques (SHAP, LIME) for transparency, and multimodal integration of Federated Learning across healthcare institutions for privacy-preserving collaborative training. Furthermore, comparisons with the deep learning models (VGG16 and DenseNet121) should be considered only as a baseline to which performance is compared, as these models are developed for large scale image data and may not be best suited for the small, structured tabular Wisconsin data. In terms of deployment, the proposed ensemble is lightweight, tabular in terms of features (nine features), meaning that CPU based inference can be performed in existing CDSS without the need for GPUs, which is also future work. Finally, ethical considerations for deployment are relevant. The data set has no patient identifiers and is publicly available and fully anonymized; however, clinical translation would need strict data governance including secure storage and controlled access to the data, with HIPAA/GDPR compliance. The model is not meant to be a self-contained diagnostic tool, but rather a tool to support clinicians in making decisions, and the need to monitor for fairness and reliability in patient populations, is ongoing.

5 Conclusion

By addressing the intrinsic shortcomings of single machine and deep learning models, this work presents a robust, hyperparameter-tuned ensemble learning system designed for accurate breast cancer classification. The proposed model achieved the highest testing accuracy on the benchmark Wisconsin Breast Cancer dataset, reaching 99.25% accuracy by combining RF, SVM, XGBoost, and a Bagging classifier through a soft-voting system and selecting their hyperparameters via the OPTUNA Bayesian framework. This was significantly better than the diagnostic accuracy of the best single traditional (SVM at 96.00%) and deep learning (VGG16 at 96.57%) baselines. The most significant clinical achievement is the ability of this optimized ensemble to achieve a recall of 1.00 for the malignant class, which substantially reduces life-threatening false-negative predictions and provides a computationally consistent paradigm well suited for deployment in modern Clinical Decision Support Systems (CDSS). Future work will focus on generalizing this framework to multimodal clinical data (e.g., raw radiological imaging) and testing its clinical applicability in more demographically heterogeneous and large-scale patient populations.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: M.M, A.I, AW, Z.M, M.T have performed conceptualization, investigation, analysis, visualization, project management, resources, editing and review. Z.M, A.W, J.A, M.T, M.A.J contribute to the software, method, data analysis, simulation, writing original draft, reviewing and editing. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available at https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original.

Ethics Approval: Not Applicable

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study

References

1. Bisoyi P. Malignant tumors–as cancer. In: Understanding Cancer. Cambridge, MA, USA: Academic Press; 2022:21–36. doi:10.1016/b978-0-323-99883-3.00011-1.
2. Breast cancer [Internet]. [cited 2026 Jan 1]. Available from: https://www.who.int/news-room/fact-sheets/detail/breast-cancer.
3. Shaikh K, Krishnan S, Thanki R. An introduction to breast cancer. In: Artificial Intelligence in Breast Cancer Early Detection and Diagnosis. Cham, Switzerland: Springer; 2020:1–20. doi:10.1007/978-3-030-59208-0_1.
4. Sriharikrishnaa S, Suresh PS, Prasada K S. An introduction to fundamentals of cancer biology. In: Optical Polarimetric Modalities for Biomedical Research. Cham, Switzerland: Springer; 2023:307–30. doi:10.1007/978-3-031-31852-8_11.
5. Harmer V. Signs and symptoms of breast cancer: the practice nurse role. *Pract Nurs*. 2016;27(8):377–82. doi:10.12968/pnur.2016.27.8.377.
6. Wang J, Wu SG. Breast cancer: an overview of current therapeutic strategies, challenge, and perspectives. *Breast Cancer Targets Ther*. 2023;15:721–30. doi:10.2147/bctt.s432526.
7. Nounou MI, ElAmrawy F, Ahmed N, Abdelraouf K, Goda S, Syed-Sha-Qhattal H. Breast cancer: conventional diagnosis and treatment modalities and recent patents and technologies. *Breast Cancer*. 2015;9s2:BCBCR-S29420. doi:10.4137/bcbcr.s29420.
8. Lakshmi D, Gurrela SR, Kuncharam M. A comparative study on breast cancer tissues using conventional and modern machine learning models. In: Smart Computing Techniques and Applications. Singapore: Springer; 2021:693–9. doi:10.1007/978-981-16-0878-0_67.
9. El Ogri O, EL-Mekkaoui J, Hjouji A. A computer-assisted medical diagnosis system for cancer diseases based on quaternion orthogonal Rademacher-Fourier moments and deep learning. *Biomed Signal Process Control*. 2026;112:108744. doi:10.1016/j.bspc.2025.108744.
10. El Ogri O, EL-Mekkaoui J, Benslimane M, Hjouji A. Automatic lip-reading classification using deep learning approaches and optimized quaternion meixner moments by GWO algorithm. *Knowl Based Syst*. 2024;304:112430. doi:10.1016/j.knosys.2024.112430.
11. Rehman A, Mahmood T, Saba T. Robust kidney carcinoma prognosis and characterization using Swin-ViT and DeepLabV^3+ with multi-model transfer learning. *Appl Soft Comput*. 2025;170:112518. doi:10.1016/j.asoc.2024.112518.
12. Al Mamun A, Bhuiyan T, Hassan MM, Anik SI. Exploring the best machine learning models for breast cancer prediction in Wisconsin. *Int J Adv Comput Sci Appl*. 2025;16(1):1362-8. doi:10.14569/ijacsa.2025.01601129.
13. Kadhim RR, Kamil MY. *Comparison of breast cancer classification models on Wisconsin dataset. Int J Reconfigurable Embed Syst*. 2022;11(2):166. doi:10.11591/ijres.v11.i2.pp166-174.
14. Shahnaz C, Hossain J, Fattah SA, Ghosh S, Khan AI. Efficient approaches for accuracy improvement of breast cancer classification using Wisconsin database. In: 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC). December 21-23, 2017. Dhaka, Bangladesh. p. 792–797. doi:10.1109/r10-htc.2017.8289075.
15. Mahmood T, Li J, Pei Y, Akhtar F, Imran A, Rehman KU. A brief survey on breast cancer diagnostic with deep learning schemes using multi-image modalities. *IEEE Access*. 2020;8:165779–809. doi:10.1109/access.2020.3021343.
16. Nissar I, Alam S, Masood S, Kashif M. MOB-CBAM: a dual-channel attention-based deep learning generalizable model for breast cancer molecular subtypes prediction using mammograms. *Comput Methods Programs Biomed*. 2024;248:108121. doi:10.1016/j.cmpb.2024.108121.
17. Ibeni WNLWH, Salikon MZM, Mustapha A, Daud SA, Salleh MNM. Comparative analysis on Bayesian classification for breast cancer problem. *Bull Electr Eng Inform.* 2019;8(4):1303–11. doi:10.11591/eei.v8i4.1628.
18. Hossin MM, Javed Mehedi Shamrat FM, Bhuiyan MR, Akter Hira R, Khan T, Molla S. Breast cancer detection: an effective comparison of different machine learning algorithms on the Wisconsin dataset. *Bull Electr Eng Inform*. 2023;12(4):2446–56. doi:10.11591/eei.v12i4.4448.
19. Mushtaq Z, Yaqub A, Sani S, Khalid A. Effective K-nearest neighbor classifications for Wisconsin breast cancer data sets. *J Chin Inst Eng*. 2020;43(1):80–92. doi:10.1080/02533839.2019.1676658.
20. Abdel-Zaher AM, Eldeib AM. Breast cancer classification using deep belief networks. *Expert Syst Appl*. 2016;46:139–44. doi:10.1016/j.eswa.2015.10.015.

21. Nissar I, Alam S, Masood S. Breast cancer molecular subtype detection through mammograms with machine learning: a comprehensive framework using radiomics and metaheuristic optimization. *Procedia Comput Sci*. 2025;258:3211–20. doi:10.1016/j.procs.2025.04.579.
22. Ahmed MR, Rahman H, Limon ZH, Siddiqui MIH, Khan MA, Pranta ASUK, et al. Hierarchical swin transformer ensemble with explainable AI for robust and decentralized breast cancer diagnosis. *Bioengineering*. 2025;12(6):651. doi:10.3390/bioengineering12060651.
23. Munshi RM, Cascone L, Alturki N, Saidani O, Alshardan A, Umer M. A novel approach for breast cancer detection using optimized ensemble learning framework and XAI. *Image Vis Comput*. 2024;142:104910. doi:10.1016/j.imavis.2024.104910.
24. Wolberg W. Breast Cancer Wisconsin (Original). UCI Machine Learning Repository. 1992.https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original
25. Rovshenov A, Peker S. Performance comparison of different machine learning techniques for early prediction of breast cancer using Wisconsin breast cancer dataset. In: 2022 3rd International Informatics and Software Engineering Conference (IISEC). December 15-16, 2022. Ankara, Turkey. p. 1–6. doi:10.1109/iisec56263.2022.9998248.
26. Jony AI, Arnob AKB. Deep learning paradigms for breast cancer diagnosis: a comparative study on Wisconsin diagnostic dataset. *Malaysian J Sci Adv Tech*. 2024:109–17. doi:10.56532/mjsat.v4i2.245.