E-ISSN 2586-9027
Homepage : https://tci-thaijo.org/index.php/SciTechAsia Science & Technology Asia
Vol.31 No.1 January - March 2026 Page: [125-144]
Original research article

Hybrid and Ensemble Learning Approaches for Accurate Breast Cancer Detection and Classification

Sridhar Siripurapu[1,*], Sateesh Gudla[2], Sridhar Bokka[3], Ramarao Bonula[4], Nataraj Dasari[5]

[1]*Nadimpalli Satyanarayana Raju Institute of Technology, Andhra Pradesh 530059, India*
[2]*Anil Neerukonda Institute of Technology and Sciences, Andhra Pradesh 531162, India*
[3]*Lendi Institute of Engineering & Technology, Andhra Pradesh 535005, India*
[4]*Aditya Institute of Technology and Management, Andhra Pradesh 532203, India*
[5]*Swarnandhra College of Engineering and Technology, Andhra Pradesh 534275, India*

Received 12 April 2025; Received in revised form 8 January 2026
Accepted 28 January 2026; Available online 27 March 2026

ABSTRACT

Breast cancer is a leading cause of mortality among women worldwide, underscoring the need for accurate differentiation between malignant and benign tumors to support early diagnosis and timely treatment. Malignant tumors are invasive and often require aggressive therapy, while benign tumors are non-cancerous and localized. Advances in Machine Learning (ML) and Deep Learning (DL) have significantly enhanced diagnostic performance in healthcare. This study explores hybrid and ensemble learning approaches—including Bagging, Boosting (AdaBoost, Gradient Boosting, and XGBoost), and Stacking—combined with traditional ML classifiers such as Logistic Regression, Support Vector Machine, K-Nearest Neighbor, Decision Tree, and Random Forest, alongside DL models including Convolutional Neural Networks and Long Short-Term Memory networks for breast cancer detection. Experiments were conducted on the Wisconsin Diagnostic Breast Cancer dataset. The workflow incorporated preprocessing, feature selection, and SMOTE-based class imbalance handling applied to training folds only. Model robustness was ensured through 10-fold stratified cross-validation with GridSearchCV hyperparameter tuning. Results show that ensemble and hybrid models outperform individual classifiers, with SVM-KNN and boosting methods demonstrating particularly strong and clinically relevant performance.

Keywords: AGBoost; Breast cancer diagnosis; Ensemble techniques; Stacking classifier; XGBoost

***Corresponding author:** sridhar.vskp@gmail.com **doi:** xx.xxxxx/scitechasia.xxxx.xx

1. Introduction and Literature Review

Breast cancer is a prevalent malignancy amongst women worldwide, marked by the uncontrolled proliferation of abnormal breast cells. Breast cancer can be categorized into invasive and non-invasive types based on its potential to spread. However, early detection through mammography, biopsy, and advanced AI-based diagnostic tools significantly improves survival rates. Advances in machine learning (ML) and deep learning (DL) significantly enhanced the computer-aided diagnosis (CAD) systems by enabling automated feature extraction, improved classification accuracy, and robust decision support for clinicians. Accordingly, plenty of research works explored the application of ML, DL and hybrid techniques for breast cancer detection and diagnosis.

Although there are numerous research works and investigations for breast cancer diagnosis, only a subset of representative works is discussed here to contextualize the proposed research. Early investigations employed classical ML and hybrid approaches like Genetic Algorithms (GA) with Artificial Neural Networks (ANN) demonstrating the feasibility of intelligent systems for breast cancer diagnosis [1]. Subsequent investigations extended these efforts to the real-time imaging environments - focusing on tracking and prediction of breast cancer to support dynamic diagnostic applications [2]. The Wisconsin Diagnostic Breast Cancer (WBCD) dataset is broadly adopted for benchmarking, with several studies reporting high diagnostic accuracy in breast cancer prediction using diversified ML models [3].

Computer-aided diagnosis (CAD) systems exploring mammographic datasets like mini-MIAS have been extensively examined – achieving high sensitivity and overall accuracy in breast cancer detection [4]. Conventional ML classifiers embracing DT and ANN models produced reliable performance with accuracies exceeding 94% [5]. Instance based classification models like KNN achieved competitive accuracy; however, subsequent adoption of CNN architectures enhanced accuracy on the same breast cancer dataset [6].

With the advances in deep learning, CNN based frameworks gained prominence for pathological breast cancer diagnosis due to strong feature extraction capability [7]. Recent studies have enhanced CNN based breast cancer diagnostic models by incorporating bio-inspired optimization techniques – alongside pretrained architectures like VGG16, ResNet-50 and DenseNet201[8]. Furthermore, hybrid CAD frameworks integrating CNN features with morphological attributes have demonstrated improved classification performance [9]. Beyond the imaging-based analysis, temporal modelling approaches employing LSTM networks in conjunction with ultra-wideband microwave signal acquisition and tumor growth simulations were proposed for reliable tumor size prediction and monitoring [10].

Most of the recent investigations emphasized hybrid and ensemble frameworks integrating DL and ML models. These include combinations of CNN, LSTM, and RF classifiers for breast cancer detection – achieving high accuracy, sensitivity and F1-scores, outperforming traditional standalone models [11]. Hybrid DL and gradient boosting methods applied to histopathological datasets like BreaKHis demonstrated effective performance in binary and multi-class classification tasks [12]. Feature extraction using DL models followed by classification with XG-

Boost was proven effective for classifying breast cancer histopathological image analysis, yielding competitive accuracy levels - surpassing the performances of previously reported methods [13].

Recent investigations (2022–2025) have increasingly explored the automated machine learning (AutoML) frameworks,

transformer-based architectures and tabular deep learning models for analysing structured biomedical data. Transformer-driven approaches, like FT-Transformer and TabTransformer employ self-attention mechanisms to effectively capture complex feature dependencies and have reported strong performance on tabular benchmarks [14-15]. Additionally, tabular architectures akin to TabNet have attracted considerable attention owing to their ability to integrate feature selection and explainability – making it attractive for clinical applications [15]. In parallel, the AutoML frameworks have advanced substantially - offering end-to-end pipelines for data preprocessing, model selection, and hyperparameter optimization – thereby achieving competitive performance with reduced human interventions [16-18].

Despite these advancements, current studies indicate that transformer-based and AutoML approaches do not constantly outperform the classical and (or) ensemble models when applied to small and well-curated noise free clinical datasets like the WDBC. These methods often need extensive tuning, larger validation cohorts, and increased computational resources to achieve effective generalization [17, 18].

Motivated by these facts, this research work is focused upon a systematic evaluation of hybrid and ensemble learning approaches for breast cancer detection - including bagging, boosting (AdaBoost, Gradient Boosting, and XGBoost), stacking, and a proposed SVM–KNN ensemble models alongside the conventional baseline classifiers. All the models were evaluated under a unified experimental framework encompassing data pre-processing, SMOTE (to reduce the class imbalances) and Ten-fold stratified cross-validation to ensure fair comparison - with emphasis on clinically relevant performance metrics. This work elucidates the practical trade-offs among deep tabular models and ensemble learners – demonstrating that carefully designed hybrid ensembles offer stable and reliable performance for small diagnostic datasets [19].

The remainder of the manuscript is organized as follows: Section-2 provides information about datasets, pre-processing steps, handling class imbalances, hybrid and ensemble methods adopted – along with the overall experimental workflow. Section-3 presents the experimental setup, performance evaluation and comparative analysis of baseline and ensemble models. Section-4 concludes the study – outlining the key findings and directions for future research.

2. Materials and Methods

2.1 Datasets information

This work utilised Wisconsin Diagnostic Breast Cancer (WDBC) dataset sourced from the Kaggle repository, which contains around 569 instances with 32 attributes. The key components are as follows: (1) ID – a unique identifier for each record; (2) Diagnosis – indicates the tumor type, where 'M' represents malignant and 'B' represents benign; (3) Mean Features – average values of various tumor characteristics, including radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension (e.g., radius_mean, texture (4) Standard Error Features – standard er-

ror values for each corresponding attribute (e.g., radius_se, texture_se); and (5) Worst-case Features – the most extreme values recorded for each trait (e.g., radius_worst, texture_worst). These features provide a comprehensive statistical representation of tumor samples - critical for accurate classification and analysis.

2.2 Workflow methodology

It is composed of the following operations:

I Data pre-processing operations

II Split the data into training data and test data.

III Perform 10-Fold Cross-Validation.

V Apply SMOTE (Synthetic Minority Over-sampling Technique) on training data.

IV Train the ensemble & baseline models.

VI Hyperparameter optimization via GridsearchCV.

VII Evaluation with test data.

VIII Perform comparative analysis.

2.3 Data pre-processing

It involves data cleaning, outliers' removal and data transformation operations (outlined below).

i) Cleaning the data ⟶ correcting the noise and inconsistencies, with missing values handled via deletion or imputation. Deletion removes unprocessable cases but is often deemed unethical in medical datasets due to data loss. Imputation replaces missing values with estimates to preserve dataset integrity - ensuring accurate disease predictions and avoiding misclassifications.

ii) Removal of Outliers ⟶ Outliers are irregular data points that impact the model performance [14]. They are commonly identified using boxplots (that display the key values - min, Q1, median Q3, max). The points outside this range are classified as outliers.

iii) Transformation of data ⟶ this process transforms data via aggregation, standardization, normalization and smoothing [15]. While normalization rescales numeric values within the 0–1 range.

2.4 K-Fold cross-validation

KCV technique [16] is widely used for model building processes to remove the dataset bias (with k=10) to achieve realistic results (Fig.1). Here the entire dataset was randomly partitioned into 10 equal-sized folds. In each iteration, one-fold is reserved for validation (testing), while the remaining nine folds are used for training the models. The entire process is repeated 10 times – ensuring that every fold serves as a validation set exactly once. Performance metrics obtained across all iterations are then aggregated to compute the final evaluation results.

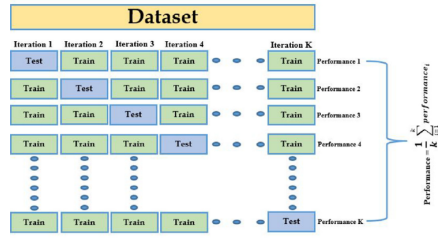

Fig. 1. K-Fold cross-validation technique.

This approach mitigates the problem

of overfitting and underfitting to achieve balanced performance on the training and testing datasets.

2.5 Synthetic Minority Over-Sampling Technique (SMOTE)

It approach is designed to address the class imbalances by producing new synthetic samples of the minority class. SMOTE creates new samples through linear interpolation of minority class samples by utilizing information from the neighboring data points - rather than merely duplicating the existing instances [17]. This augmentation increases the diversity of minority class samples and consequently enhances the learning capability and predictive performance of ML models on imbalanced datasets [18]. For a given minority class sample zi, the-k nearest neighbor (in feature space) of minority class sample is computed Eq. (1) by its Euclidean distance:

$$d(z_i, z_j) = \sqrt{\sum_{i=1}^{p} (z_{i,l}, z_{j,l})^2}. \qquad (2.1)$$

The synthetic samples are generated through the following equation:

$$Z_{synthetic} = \delta z_j + (1 - \delta) z_i. \qquad (2.2)$$

2.6 Proposed baseline AI models

2.6.1 Logistic Regression (LR)

LR estimates the class probabilities with outputs ranging between 0 and 1. Its performance is dependent on data pre-processing operations: cleaning, handling missing values, and feature selection. All the patient attributes were analyzed for correlation with target variable to determine the key predictors. LR delivers probabilities for categorical outcomes (e.g., True/False or 0/1).

2.6.2 Support Vector Machine (SVM)

Typically, SVM identifies an optimal hyperplane to distinguish between data points of various classes. A hyperplane in n-dimensional space is defined by:

$$wTx + b = 0, \qquad (2.3)$$

where wTx denotes dot product and b is the bias term. When data isn't linearly divisible, SVM employs a "kernel trick" to shift it into a higher-dimensional realm where a linear split becomes feasible.

2.6.3 K-Nearest Neighbor (KNN)

KNN is a distance-based learning algorithm used for classification and regression in which predictions happen by identifying the nearest data points based on the Euclidean distance.

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + \cdots + (p_n - q_n)^2}. \qquad (2.4)$$

In classification tasks, KNN assigns a class label based on the majority class among its k-nearest neighbors, whereas in regression tasks, it predicts the output by computing the average value of those neighbors.

2.6.4 Decision Tree (DT)

DT classifier partitions data into a hierarchical structure of branches based on simple decision rules (Figure.2). Beginning at the root node, the algorithm recursively divides the data into subsets using feature-based thresholds that best separate the classes - guided by criteria such as Gini impurity or information gain. This process continues until terminal leaf nodes are reached, which represents the final class predictions.

Predictions are made by tracing a new data point through the tree to a leaf

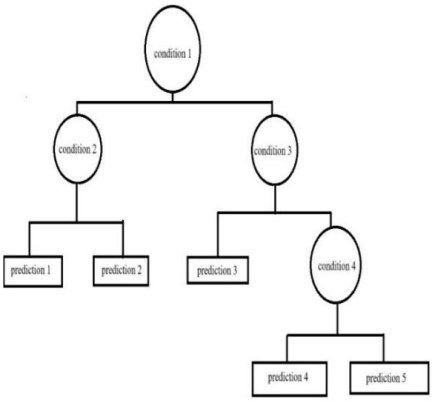

Fig. 2. Decision Tree algorithm.

node. The method is straightforward, like a flowchart, and provides clear, interpretable results.

2.6.5 Random Forest (RF)

RF algorithm performs predictions by aggregating the outputs of multiple decision trees, each trained on randomly chosen subsets of data and feature space. For classification, it uses majority voting; for regression, it takes the average. This method reduces overfitting and improves accuracy compared to a single decision tree.

2.6.6 Convolutional Neural Network (CNN)

CNN is a specialized deep learning model primarily designed for image processing (Fig. 3); it consists of three main types of layers:

i) Convolutional Layer – extracts spatial features and reduces dimensionality with help of pooling layers.

ii) Pooling Layer – reduces spatial dimensions to lower computational cost and prevents overfitting.

iii) Fully Connected Layer – processes the extracted features and makes final predictions.

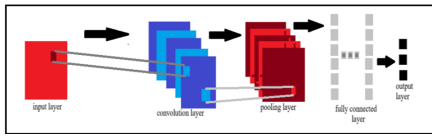

Fig. 3. Structure of Convolutional Neural Network.

Images are represented as a 3D matrix of pixel values. To introduce non-linearity, the ReLU function is applied:

$$f(z) = \max(0, z). \tag{2.5}$$

After multiple convolutional and pooling layers, the feature maps are:

$$F = Flatten(P). \tag{2.6}$$

A fully connected layer applies:

$$y = f(WF + b), \tag{2.7}$$

where W is weight matrix, F is input feature vector, b is bias term, f is activation function.

2.6.7 Long Short-Term Memory (LSTM)

LSTM is used to capture long-term dependencies, making it effective for sequential data like time-series or speech recognition. LSTM cell (Fig. 4) is comprised of the following list.

i) Forget Gate decides which information to discard.

$$f_t = \sigma(Wf[h_t - 1, x_t] + b_f), \tag{2.8}$$

where f_t is forget gate activation, W_f is forget gate weight matrix, b_f is forget gate bias σ is sigmoid activation function (output range: 0 to 1).

ii) Input Gate determines which new information to store.

$$i_t = \sigma(W_i[h_t - 1, x_t] + b_i). \tag{2.9}$$

where i_t is input gate activation; W_i, W_c are weight matrices.

iii) Cell State stores long-term memory.

$$C_t = \tanh(W_C[h_t - 1, x_t] + b_c), \quad (2.10)$$

where Ct is Candidate cell state; tanh is Hyperbolic tangent activation.

iv) Output gate controls what the output is as follows.

$$o_t = \sigma(W_o[h_t - 1, x_t] + b_o), \quad (2.11)$$

where o_t is output gate activation, W_o is output gate weight matrix, b_o is bias.

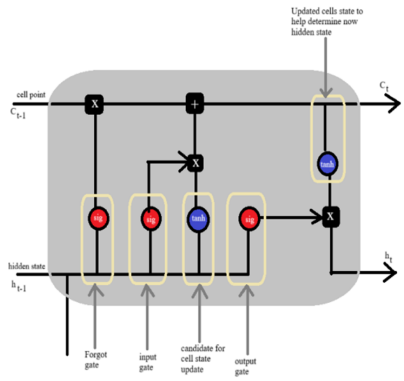

Fig. 4. LSTM Unit Cell.

2.7 Hybrid and ensemble models

2.7.1 AdaBoost (Adaptive Boosting)

AdaBoost is an ensemble boosting algorithm that enhances classification performance by sequentially emphasizing misclassified samples through adaptive weight adjustments (Fig. 5) - such that model focuses more on difficult cases in the next iteration. The training dataset comprises input feature vectors and their corresponding class labels.

Each training sample is initially given equal weight of

$$D_t(i) = \frac{1}{N}, \ \forall i = 1, ..., N, \quad (2.12)$$

where, $D_t(i)$ is weight distribution.

Now training the weak classifier:

$$\epsilon_t = \sum_{i=1}^{N} D_t(i) I(y_i \neq h_t(x_i)). \quad (2.13)$$

The weak classifier weight is now computed. Then sample weights are updated by increasing the weights of misclassified points and decreasing weights of correctly classified points. The final classifier is a weighted sum of weak classifiers:

$$H(x) = \text{sign}(\sum_{t=1}^{T} \alpha_t h_t(x)), \quad (2.14)$$

where sign determines the final class +1.

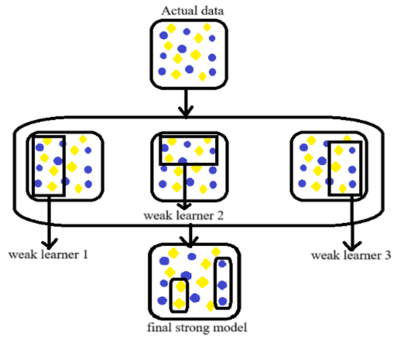

Fig. 5. AdaBoost Block Diagram.

2.7.2 Extreme Gradient Boosting

XGBoost commences with an input dataset (root node) and the boosting process is carried on with multiple decision trees.

In the flowchart (Fig. 6), yellow circles represent the decision nodes, while green circles indicate leaf nodes, Blue arrows depict the sequential enhancement of trees across boosting iterations, and red arrows trace the corresponding decision

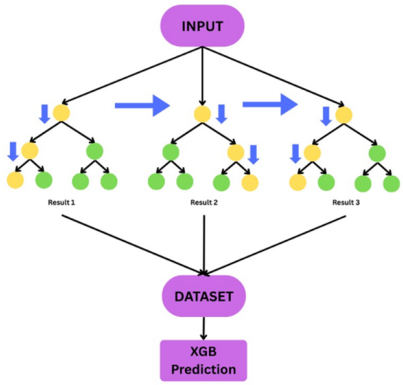

Fig. 6. XGBoost Block Diagram.

paths. Multiple trees labeled as Result-1, Result-2, and Result-3- denote various boosting rounds. The outputs of all trees are combined to form a single dataset. Final prediction is obtained by aggregating results from all trees.

2.8 Hyperparameter Tuning

Typically, Hyperparameters influence the model's learning process. While poorly tuned hyperparameters lead to under-performance and inaccurate predictions. Hyperparameter tuning [19] ensures better learning of data to avoid overfitting or underfitting and ultimately arrive at accurate predictions of breast cancer. However, tuning parameters like regularization strength and kernel type enhances accuracy for the adopted models.

2.9 Performance metrics

The performance assessment metrics: Accuracy, Precision, Recall, F1-Score and Specificity are briefly outlined below.

i) Accuracy: denotes the ratio of correct predictions to the total number of predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{2.15}$$

ii) Precision: indicates the proportion of correctly predicted positive cases. It is helpful in cases where False Positives are prioritized than False Negatives.

$$\text{Precision} = \frac{\text{Correct Prediction}}{\text{Total Predictions}} = \frac{TP}{TP + FP}. \tag{2.16}$$

iii) Recall (Sensitivity or True Positive Rate): refers to the proportion of correctly identified actual positive cases. It is vital in applications where the number of false negatives outweigh the number of false positives.

$$\text{Recall} = \frac{\text{Correct Prediction}}{\text{Total Ground Truth}} = \frac{TP}{TP + FN}. \tag{2.17}$$

iv) F1-Score: depicts harmonic mean among precision and recall values-integrating them into a single value.

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{TP}{TP + FN}. \tag{2.18}$$

v) Specificity: determines the model's ability to appropriately compute the actual negative cases.

$$\text{Specificity} = \frac{\text{Actual Negatives}}{\text{Total Ground Truth}} = \frac{TN}{TN + FP}. \tag{2.19}$$

3. Results and Discussions

All the experiments were conducted within the Jupyter Notebook environment, employing Scikit-Learn, NumPy, Pandas, and

Matplotlib for data preprocessing, model training, and performance evaluation. The WDBC dataset was subjected to

multiple pre-processing operations. The entire dataset was split into training and testing subsets using an 80:20 split ratio. Additionally 10-fold cross-validation (KCV) and SMOTE were applied to mitigate the class imbalances. The resultant accuracy values are presented in Table 1. Both ensemble models and baseline models were trained after hyperparameter optimization via GridSearchCV. Evaluations were based on key clinical metrics like Accuracy, Precision, Recall, F1-score, AUC-ROC, and Precision–Recall curves.

3.1 Heat Map

illustrates the variable correlations (Fig. 7), where brighter colors (like yellow and orange) denote strong positive relationships, while darker shades (like black or deep red) indicate weak or negative correlations.

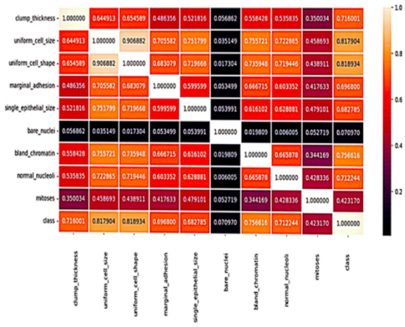

Fig. 7. Validation of dataset (heat map).

Key features like uniform_cell_size, uniform_cell_shape, bland_chromatin, normal_nucleoli, and clump_thickness exhibit strong correlations - highlighting their importance in prédictive modeling. In contrast, bare_nuclei shows weaker correlation – demonstrating minimal influence. The heatmap reveals no significant negative corrélations, while high correlation values (>0.8) suggest feature redundancy and multicollinearity thereby affecting the model stability. These issues can be mitigated through appropriate feature selection or regularization techniques to enhance the overall model performance.

3.2 Analysis of Confusion Matrix (CM) and Receiver Operating Characteristic (ROC) Curves

Typically, CM evaluates classification performance through comparison of predicted labels with actual outcomes. In binary classification, CM includes True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) - While, the ROC curve depicts a trade-off between True Positive and False Positive Rates. On the other hand, Area Under Curve (AUC) serves as a key performance metric–such that an AUC of 1.0 implies the distinction of various classes flawlessly. Experimental evaluations depicting the ensemble models and other ML classifiers along with their comparative analysis are discussed in this section of the manuscript.

3.3 Evaluation of Advanced Learning Models

Typically, CM evaluates classification performance through comparison of predicted labels with actual outcomes. In binary classification, CM includes True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) , while, the ROC curve depicts a trade-off between True Positive and False Positive Rates. On the other hand, Area under Curve (AUC) serves as a key performance metric. AUC of 1.0 implies the distinction of various classes flawlessly. Experimental evaluations of ensemble models and other ML classifiers are discussed in this section of the manuscript.

3.3.1 CNN Mode

CM of CNN model (Fig. 8) demonstrated strong performance by correctly identifying 35 out of 36 (high true positive rate) malignant cases.

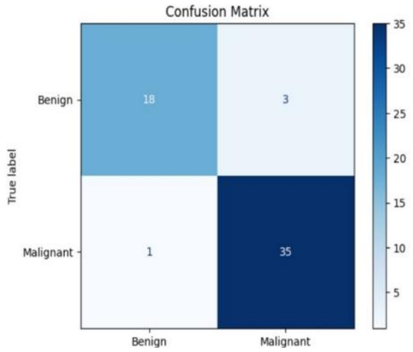

Fig. 8. Confusion Matrix of CNN Model.

For benign cases, the model achieved a slightly lower accuracy – correctly classifying 18 out of 21 instances. Fig. 9 illustrates the ROC and training behavior of CNN model – demonstrating low values of training and validation loss between the epochs 15 and 20, thereby indicating the effective learning ability of the model. However, the training loss continued to decline beyond this range while the validation loss plateaus – suggesting the onset of overfitting.

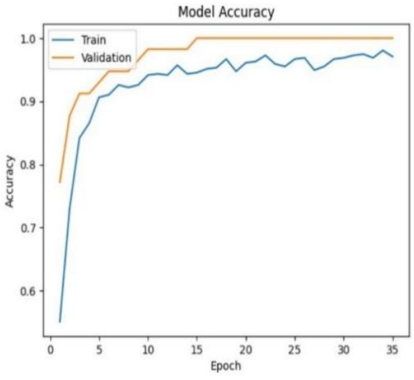

Fig. 9. ROC of Training, Validation Loss.

The ROC plot illustrates the training and validation accuracies, where the blue curve representing training accuracy approached 100% with a steady rise. In contrast the validation curve (in orange) demonstrated an improvement but leveled off earlier - indicating limited generalization beyond a certain point.

3.3.2 LSTM Model

CM of LSTM Model (Fig. 10) classified 41 malignant cases and 70 benign cases – demonstrating strong predictive accuracy.

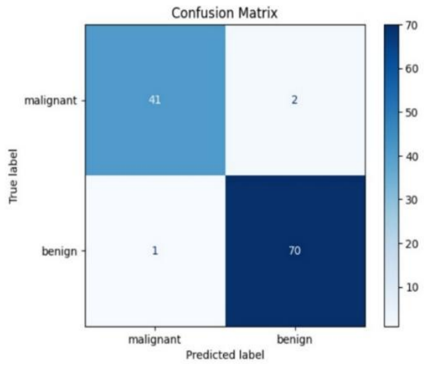

Fig. 10. Confusion matrix of LSTM.

Only 2 malignant cases were misclassified as benign. There was a single false positive case – illustrating an incorrect classification of a benign case as malignant. Based on these results - the model exhibited high sensitivity and specificity, indicating strong reliability in identifying benign cases while maintaining a low error rate in the detection of malignancies.

The ROC plot of LSTM (Fig.11) demonstrated an AUC value of 1.00 – indicating 100% sensitivity (true positive rate) with a 0 false positive rate in most cases. While the diagonal red line depicted the random classifier (AUC = 0.5). A significant deviation of blue curve above the red line demonstrated a strong predictive capability of the model. In overall, LSTM model ex-

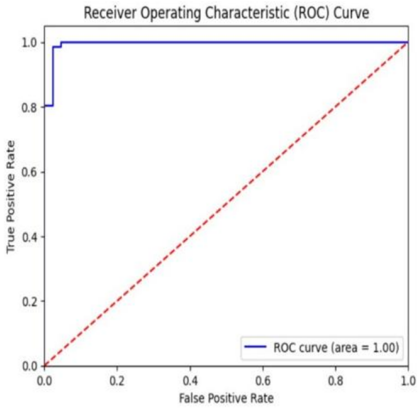

Fig. 11. ROC of LSTM.

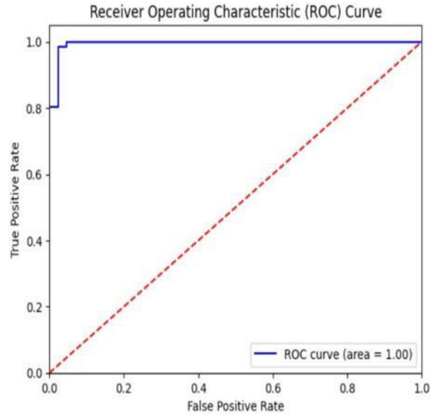

Fig. 13. ROC of AdaBoos.

hibited exceptional accuracy in the detection of breast cancer.

3.3.3 AdaBoost Mode

CM of AdaBoost model (Fig.12) correctly classified 69 benign cases and misclassified 2 benign cases as malignant (false positives).

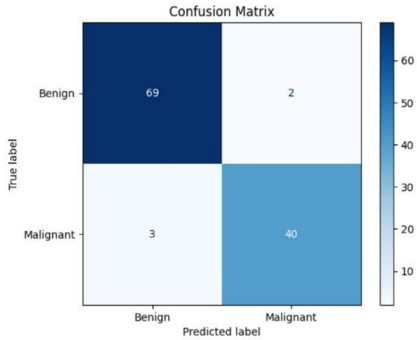

Fig. 12. Confusion matrix of AdaBoos.

While 40 malignant cases were correctly classified, 3 malignant cases were misclassified as benign (false negatives). Overall, the model demonstrated superior accuracy and reliability.

The ROC plot of AdaBoost model (Fig. 13) depicted AUC of 1.00 – indicating a perfect distinction between the two classes. The red dashed line depicted a random classifier (AUC = 0.5), and the model's ROC curve lying well above this line confirmed its high predictive power.

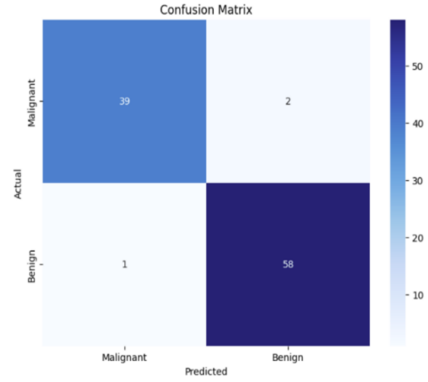

Fig. 14. Confusion matrix of XGBoost.

Results obtained clearly depicted the reliability of XGBoost; it misclassified 2 malignant cases as benign - resulting in false negatives. ROC plot (Fig.15) demonstrated high sensitivity with a low false positive rate. However, the AUC value (of 1.00) signified exceptional classification ability of XGBoost in effectively distinguishing the malignant and benign cases with absolute accuracy - thereby minimizing the false negatives and false positives.

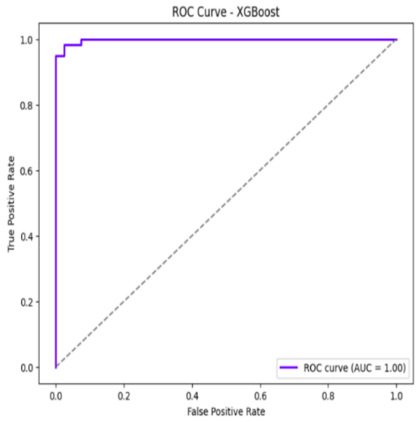

Fig. 15. ROC of XGBoost.

3.4 Comparative Analysis of Advanced Learning Models

Table.1 shows a comparative analysis of CNN, LSTM, AdaBoost and XGBoost – across the four key performance metrics. It was observed that XGBoost achieved highest accuracy of 97.71%, demonstrating a strong balance between precision and recall, particularly in identifying the malignant cases.

Table 1. Comparative Analysis.

Model	Accuracy	Precision	Recall	F1_score
CNN	91.21%	96.97%	88.80%	92.75%
LSTM	97.00%	88.10%	92.11%	91.44%
AdaBoost	96.93%	96.41%	85.38%	83.70%
XGBoost	97.71%	96.66%	98.30%	97.47%

LSTM followed closely with an accuracy of 97.00%, exhibiting stable performance across all metrics. AdaBoost also attained high accuracy of 96.93% - however, its comparatively lower recall (85.38%) and F1-score (83.70%) indicated limitations in detecting all malignant cases.

CNN demonstrated moderate performance with an accuracy of 91.21% and recall of 88.80% - reflecting a more conservative prediction behavior, although its F1-score (92.75%) remained competitive. Confusion matrix analysis revealed that all models performed reliably well with minor variances. LSTM produced the minimum misclassifications (with only 2 false positives and 1 false negative) making it particularly effective in correctly identifying benign cases. AdaBoost showed a larger number of false negatives, occasionally misclassifying malignant tumors, while XGBoost exhibited slightly fewer true negatives – indicating a tendency to misclassify few benign cases as malignant.

ROC analysis further demonstrated that XGBoost, LSTM and AdaBoost attained perfect discrimination with an AUC of 1.00 – confirming their strong capability in distinguishing the benign and malignant tumors. In contrast, CNN demonstrated a lower AUC (0.91) depicting a comparatively weaker separability.

3.5 Evaluation of ensemble models

Upon the evaluation of baseline models (LR, SVM, KNN and DT) – different hybrid and ensemble models such as Bagging, AdaBoost, Gradient Boosting, XGBoost, Stacking and SVM-KNN were evaluated under similar cross- validation settings to ensure fair comparison.

3.5.1 SVM and KNN ensemble

In the proposed hybrid framework, SVM classifier directly classifies points that are linearly separable, while KNN is employed to handle ambiguous instances by analyzing their local neighborhood. Those uncertain samples that lie close to the SVM decision boundary are delegated to the KNN for refined classification. This collaborative strategy enhances the predictive accuracy in overlapping class regions, where SVM puts much effort. Moreover, computational complexity is reduced by assigning straightforward classifications to SVM-while KNN effectively captures the

non-linear decision boundaries more effectively – resulting in enhanced overall performance.

CM of SVM and KNN Ensemble (Fig. 16) indicates that the model correctly classified 182 benign samples, with only 2 benign cases misclassified as malignant - demonstrating high specificity in identifying the benign cases.

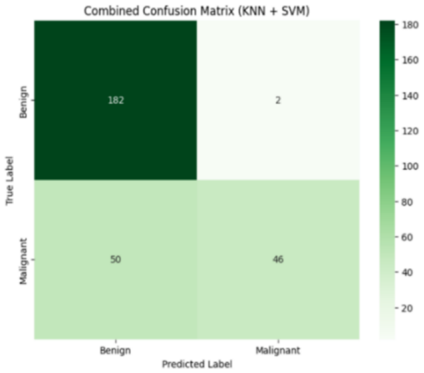

Fig. 16. CM of SVM & KNN.

However, the model correctly identified 46 malignant cases and misclassified 50 malignant samples as benign - resulting in high false-negative rate. This significant misclassification of malignant cases may sometimes lead to delayed diagnosis and treatment.

3.5.2 Bagging Classifier and Decision Tree

In the hybrid Bagging and DT ensemble model (Fig. 17). Better performance is achieved by combining multiple DT classifiers rather than relying on a single tree.

Multiple bootstrap samples are generated from the original dataset, and an independent DT is trained on each sample. The bagging classifier then integrates the predictions of all trained trees using majority voting for classification tasks.

CM of bagging and DT ensemble

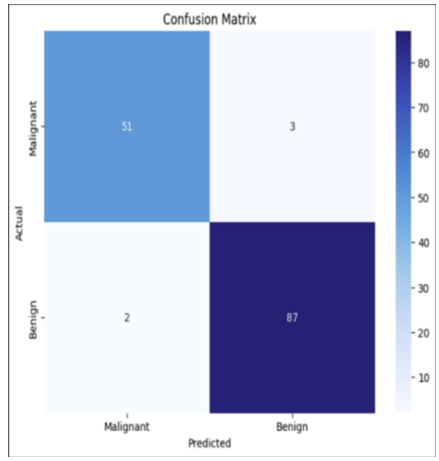

Fig. 17. CM of Bagging and DT ensemble.

model (Fig. 17) correctly classified 51 malignant cases with only 3 FNs. While 87 benign cases were accurately identified with just 2 FPs. This balanced performance reflected high sensitivity and specificity along with low misclassification rates – thereby highlighting the model's accuracy and reliability for medical diagnosis.

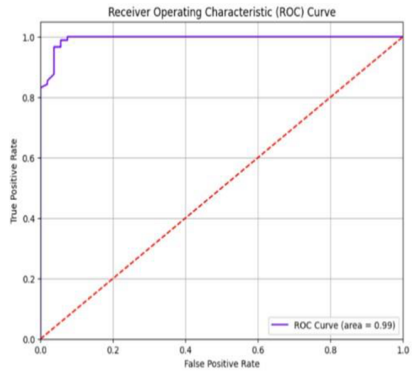

Fig. 18. ROC of Bagging and DT ensemble.

Furthermore, the ROC plot (Fig. 18) demonstrated high true positive rate accompanied by a minimal false positive rate - while AUC of 0.99 highlighted the classifier's excellent ability to distinguish the malignant and benign cases. These results indicate strong robustness and generalization

ability - confirming the model as a reliable choice for breast cancer classification.

3.5.3 AdaBoost and Gradient Boost with XGBoost

Hybrid and ensemble models represent powerful earning strategies - particularly when integrating methods like AdaBoost, Gradient Boosting, and XGBoost. AdaBoost primarily emphasizes the correction of weak learners. Gradient Boosting minimizes the residual errors through a gradient descent-based optimization framework - while XGBoost enhances the performance via regularization and parallel computing mechanisms. Stacking multiple boosting algorithms leverages their complementary strengths, whereas integration of bagging and boosting techniques improves the prediction stability and enhances the overall performance of advanced models such as XGBoost [22].

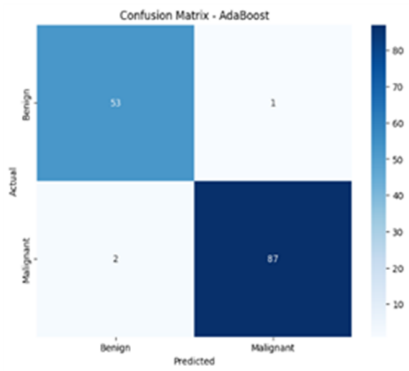

Fig. 19. CM of AdaBoost ensemble.

CM of AdaBoost classifier (Fig. 19) reported 53 TPs, 87 TNs, 1 FP, and 2 FNs – demonstrating a high classification accuracy with minimal misclassifications.

CM of Gradient Boosting classifier (Fig. 20) reports 51 TPs, 86 TNs, 3 FPs, and 3 FNs – yielding 137 correct predictions—slightly fewer than AdaBoost. The hike in false negatives literally indicates a marginal decline in the detection of malignant cases – critical for breast cancer diagnosis.

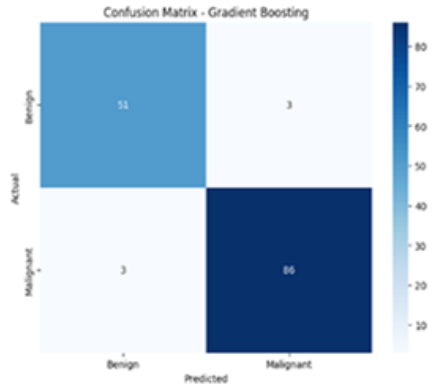

Fig. 20. CM of Gradient Boost ensemble.

CM of XGBoost classifier (Fig. 21) mirrored the performance of Gradient Boosting with identical number of TPs, TNs, FPs, and FNs – resulting in 137 correct predictions. Overall, both the XGBoost and Gradient Boosting approaches offered balanced performance.

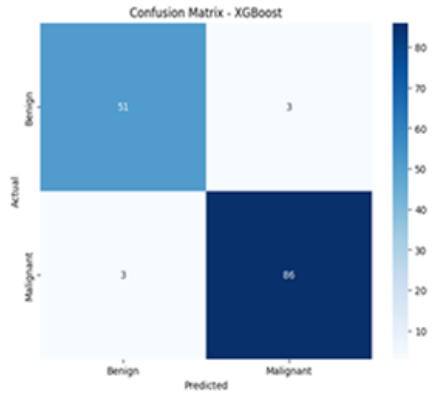

Fig. 21. CM of XGBoost ensemble.

AUC (0.98) of AdaBoost depicting high true positive rates and low false positive rates signified effective detection of malignant cases with minimal number of false positives. While AUC (0.96) of Gradient Boosting demonstrated a marginal rise in false positives for comparable sensitiv-

ity. In contrast, XGBoost secured highest AUC of 0.99, reflecting exceptional ability to distinguish the malignant and benign cases. Therefore, XGBoost is considered to effectively trade-off between sensitivity and specificity, thereby confirming it as the most reliable classifier among the three.

3.5.4 Stacking Gradient Boosting and RF

A powerful hybrid approach involves usage of Gradient Boosting and Random Forest as base models [27] whose predictions are given to a meta-classifier to generate final output. Initially, Gradient Boosting and Random Forest classifiers are trained independently. Further, their predictions are utilized as input features for a meta-classifier - which are then combined to obtain improved accuracy. Here Gradient Boosting effectively captures complex data patterns – while Random Forest mitigates overfitting through feature randomness. Meta-classifier integrates the strengths of individual models to boost the complete performance.

CM of stacking model (Fig. 22) correctly identified 69 true positives, 40 true negatives, 2 false positives, and 3 false negatives.

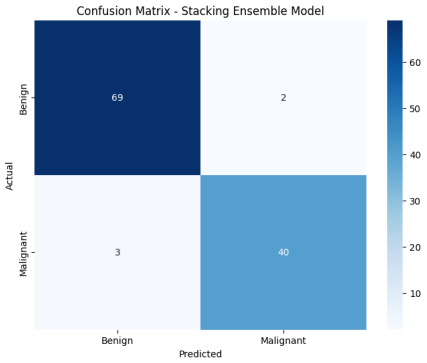

Fig. 22. CM of Stacking ensemble model .

This model demonstrated strong performance in breast cancer classification - with 109 correct predictions (out of 114 cases). The presence of 3 false negatives highlighted a critical area for improvement - because misclassification of malignant cases was highly undesirable in medical diagnostics.

The ROC plot in blue (Fig. 23) obtained impressive AUC of 0.98 - effectively distinguishing the positive and negative classes among different threshold values.

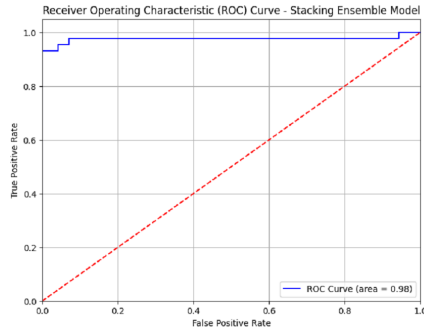

Fig. 23. ROC of Stacking ensemble model .

The plot closely reflected a high true positive rate with minimal FP rate (ideal for classification tasks). For comparison, the red-dashed line denoted performance of random classifier with AUC of 0.5, emphasizing strong predictive capability of the model.

3.6 Comparative Analysis of Baseline and ensemble models

The comparative analysis illustrated in Table.2 demonstrates that hybrid and ensemble models consistently outperform the individual baseline models under the same experimental pipeline. Performance metrics are reported as mean ± standard deviation (SD) highlighting both accuracy and stability across folds. Among the evaluated approaches, the SVM and KNN ensemble achieved highest accuracy of 99.98

Table 2. Performance Analysis of Baseline and ensemble models.

Model	Accuracy	Precision	Recall	F1-Score
Baseline Models				
LR	97.25 ± 0.23	96.38 ± 0.13	96.54 ± 0.12	96.73 ± 0.41
SVM	97.42 ± 0.11	97.38 ± 0.21	97.53 ± 0.22	97.40 ± 0.30
KNN	96.41 ± 0.31	96.58 ± 0.40	96.34 ± 0.50	96.30 ± 0.40
DT	95.56 ± 0.52	95.48 ± 0.62	95.65 ± 0.61	95.43 ± 0.63
Ensemble Models				
SVM + KNN	99.98 ± 0.08	98.01 ± 0.09	97.26 ± 0.10	95.59 ± 0.09
Bagging (DT)	97.51 ± 0.11	97.48 ± 0.12	98.55 ± 0.12	98.43 ± 0.11
AdaBoost	98.25 ± 0.12	99.65 ± 0.11	98.14 ± 0.12	98.20 ± 0.11
Gradient Boost	96.20 ± 0.11	97.68 ± 0.12	97.09 ± 0.12	97.13 ± 0.11
XGBoost	96.20 ± 0.10	97.68 ± 0.11	97.09 ± 0.11	97.13 ± 0.11
Stacking (GB + RF)	96.44 ± 0.13	95.67 ± 0.14	93.37 ± 0.15	94.38 ± 0.14

± 0.08% demonstrating a notable improvement over the baseline LR and SVM models, along with strong precision and recall values.

The AdaBoost model also exhibited competitive performance, with an accuracy 98.25±0.12% indicating high sensitivity in identifying malignant cases. In contrast, Gradient Boost and XGBoost achieved comparable accuracies of 96.20%+0.11 suggesting consistent generalization across the cross-validation folds. The stacking ensemble (Gradient Boosting + Random Forest delivered balanced performance with slightly higher variability, indicating moderate sensitivity to the fold composition. Overall, the modest, yet non-zero standard deviations across all the models confirm numerical stability - consistent with the structured and low noise nature of the WBC dataset.

Further analysis using confusion matrices revealed that the bagging-DT ensemble effectively minimized false negatives while maintaining strong detection capability for malignant cases. Although the SVM and KNN ensemble achieved highest accuracy, it exhibited relatively higher misclassification of malignant cases compared to ensemble counterparts. The Stacking ensemble demonstrated reasonable performance with a larger true negatives and comparatively fewer true positives. Overall, AdaBoost and Bagging-DT ensembles depicted consistent and reliable behavior for tumor detection.

The ROC curve analysis further substantiates these findings. The Bagging-DT ensemble achieved an AUC of 0.99, indicating excellent discriminative capability. While the SVM-KNN ensemble and AdaBoost models followed closely with AUC values of 0.98, while Gradient Boosting grasped an 0.96. Notably, XGBoost achieved a top AUC of 0.99, and the Stacking Ensemble maintained an AUC of 0.98. These results confirm that hybrid and ensemble models outperformed individual classifiers in distinguishing the benign and malignant cases - as manifested in the CM, ROC curves, and tabulated performance metrics.

Different studies using the WDBC dataset and other datasets clearly demonstrated that ensemble methods exhibit consistently better performances despite competitive performances by the baseline models.

A comparative analysis with the existing studies, summarized in Table 3, fur-

Table 3. Comparative Analysis with other works.

Reference (Year)	Method	Dataset	Accuracy
28 (2023)	XGBoost, LR, KNN, RF, DT	WBCD	83%
29 (2022)	KNN, LR, RF, DT & ANN	BC database of Coimbra (UCI)	64%
30 (2022)	SVM	MIAS Dataset of Mammograms	87.10%
31 (2022)	XGBoost	WBCD	97%
32 (2022)	GBM, XGBoost, LightGBM	WBCD	LGBM has shown robust results
33 (2021)	KNN, SVM & DT	Breast images	—
34 (2021)	SVM, KNN, LR, RF, DT & NB	WBCD	96.5%
35 (2020)	LR & DT	WBCD	95% & 94%
36 (2020)	LR, KNN, DT & NB	WBCD	LR achieved good results
37 (2020)	NE Model	WBCD	This ensemble model outperformed other models
Proposed Work	SVM & KNN, AdaBoost	WBCD	99.98%, 98.52%

ther strengthens these observations. Previous works on the WBCD and related datasets consistently report improved performance using ensemble techniques despite competitive results from baseline models. In comparison, the proposed SVM-KNN and AdaBoost ensemble models perform even better relative to prior works. Although AdaBoost and Gradient Boosting (with XGBoost) obtained accuracies of 98.25%, and 96.20% respectively – the overall findings confirm the robustness and effectiveness of hybrid and ensemble strategies for breast cancer detection.

3.7 Ethical and Clinical considerations

Though the proposed ensemble models achieved strong performances on WBC dataset, there are many factors of concern before their deployment in clinical settings. They are listed as follows:

i) Typically, the real-world breast cancer data always exhibit high levels of class imbalances, increased noise and population heterogeneity on comparison with the curated benchmark datasets – which will impact the generalization ability of the models.

ii) Therefore, validation on multi-institutional datasets is inevitably required to ensure robustness among diverse patient populations.

iii) Clinical adoption of models necessitates adequate model interpretability and explainability to support the clinician in the decision-making process.

iv) False-negative predictions pose clinical risk in breast cancer diagnosis - as skipped malignant cases might delay the treatment and have adverse patient outcomes.

4. Conclusion and Future Scope

The findings of this research work are mentioned below.

1. This study presented a systematic comparative evaluation of baseline, deep learning, and hybrid/ensemble models for breast cancer detection using the WDBC dataset under a unified framework.

2. Hybrid and ensemble approaches consistently outperform individual baseline classifiers, achieving higher accuracy, recall, stability, and diagnostic reliability.

3. The proposed SVM–KNN ensemble attains the highest accuracy (99.98%), demonstrating the benefit of combining complementary classifiers while indicating a trade-off with F1-score and class balance.

4. AdaBoost effectively minimizes false negatives, highlighting clinical suitability, while Bagging (DT), KNN, and XGBoost achieve the highest ROC–AUC values (0.99); the stacking ensemble shows balanced performance.

5. The application of SMOTE exclusively on training folds and 10-fold stratified cross-validation ensured fair evaluation and robust generalization across all models.

6. Comparative analysis with prior studies confirms that ensemble learning provides reliable and reproducible performance gains for small, well-curated diagnostic datasets.

7. Overall, the findings demonstrate that well-designed hybrid and ensemble frameworks offer robust, accurate, and clinically relevant solutions, with future work focused on real-time deployment, multimodal integration, external validation, and explainability

References

[1] Singh I, Sanwal K, Praveen S. Breast cancer detection using two-fold genetic evolution of neural network ensembles. In: 2016 International Conference on Data Science and Engineering (ICDSE). 2016.

[2] Eman N, Cabatuan MK, Dadios EP, Gan Lim LA. Detecting and tracking female breasts using neural network in real-time. In: 2013 IEEE Region 10 Conference (TENCON). 2013.

[3] Gupta N, Gupta HK, Srivastava R, Saxena C, Surjeete. Design and implementation of artificial neural network classifiers based on hypertuning parameters for breast cancer diagnosis. Procedia Comput Sci. 2024.

[4] Tariq N. Breast cancer detection using artificial neural networks. J Mol Biomark Diagn.

[5] Higa A. Diagnosis of breast cancer using decision tree and artificial neural network algorithms. Int J Comput Appl Technol Res.

[6] Wadhwa G, Mathur M. A convolutional neural network approach for the diagnosis of breast cancer. In: 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC). 2020.

[7] Liang Y, Yang J, Quan X, Zhang H. Metastatic breast cancer recognition in histopathology images using convolutional neural network with attention mechanism. In: 2019 Chinese Automation Congress (CAC). 2019.

[8] Gonçalves CB, Souza JR, Fernandes H. CNN architecture optimization using bio-inspired algorithms for breast cancer detection in infrared images. Comput Biol Med. 2021.

[9] Wang Z, Li M, Wang H, Jiang H, Yao Y, Zhang H, et al. Breast cancer detection using extreme learning machine based on feature fusion with CNN deep features. IEEE Access. 2019.

[10] Wu Y, Zhu H. Investigation of long short-term memory based ultrawide band microwave breast tumor size prediction. In: 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). 2019.

[11] Begum A, Kumar VD, Asghar J, Hemalatha D, Arulkumaran G. A combined deep CNN-LSTM with a random forest approach for breast cancer diagnosis. 2022.

[12] Liew XY, Hameed N, Clos J. An investigation of XGBoost-based algorithm for breast cancer classification. Mach Learn Appl. 2021.

[13] Maleki A, Raahemi M, Nasiri H. Breast cancer diagnosis from histopathology images using deep neural network and XGBoost. Biomed Signal Process Control. 2023.

[14] Gorishniy Y, et al. Revisiting deep learning models for tabular data. In: Advances in Neural Information Processing Systems (NeurIPS). 2021.

[15] Arik SO, Pfister T. TabNet: Attentive interpretable tabular learning. In: AAAI Conference on Artificial Intelligence. 2021.

[16] Huang H, et al. Transformer-based models for tabular data: A survey. IEEE Access. 2023.

[17] Hutter F, Kotthoff L, Vanschoren J, editors. Automated machine learning: Methods, systems, challenges. Springer; 2022.

[18] Shwartz-Ziv R, et al. Tabular deep learning: A review. ACM Comput Surv. 2024.

[19] Author(s). AutoML for clinical decision support systems. IEEE J Biomed Health Inform. 2023–5.

[20] Pranatha MDA, Pramaita N, Sudarma M, Widyantara IMO. Filtering outlier data using box whisker plot method for fuzzy time series rainfall forecasting. In: 2018 4th International Conference on Wireless and Telematics (ICWT). IEEE; 2018. p.1–4.

[21] Shah D, Patel S, Bharti SK. Heart disease prediction using machine learning techniques. SN Comput Sci. 2020;1:1–6.

[22] Nti IK, Boateng ON, Aning J. Performance of ML algorithms with different K values in KCV. Int J Inf Technol Comput Sci. 2021;6:61–71.

[23] Fernandez A, et al. SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. J Artif Intell Res. 2018;61:863–905.

[24] Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics. 2013;14:106.

[25] Bergstra J, et al. Algorithms for hyper-parameter optimization. In: Advances in Neural Information Processing Systems. 2011;24.

[26] JaiKrishna SV, Chantarakasemchit O, Meesad P. A breakup machine learning approach for breast cancer prediction. In: 2019 11th International Conference on Information Technology and Electrical Engineering (ICITEE). 2019.

[27] Dhungel N, Carneiro G, Bradley AP. Automated mass detection in mammograms using cascaded deep learning and random forests. In: 2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA). 2015.

[28] Botlagunta M, et al. Classification and diagnostic prediction of breast cancer metastasis on clinical data using machine learning algorithms. Sci Rep. 2023;13(1):1–17.

[29] Binsaif N. Application of machine learning models to the detection of breast cancer. Mobile Inf Syst. 2022;2022.

[30] Al-Fahaidy FAK, Al-Fuhaidi B, Al-Darouby I, Al-Abady F, Al-Qadry M, Al-Gamal A. A diagnostic model of breast cancer based on digital mammogram images using machine learning techniques. Appl Comput Intell Soft Comput. 2022;2022:1–17.

[31] Ghosh P. Breast cancer Wisconsin (diagnostic) dataset. UCI Machine Learning Repository; 2022. Available from: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)

[32] Akbulut S, Cicek IB, Colak C. Classification of breast cancer on the strength of potential risk factors with boosting models: A public health informatics application. Med Bull Haseki. 2022;60:196–203.

[33] Dey N, Rajinikanth V, Hassanien AE. An examination system to classify breast thermal images into early/acute DCIS class. In: International Conference on Data Science and Applications. Springer; 2021. p.209–20.

[34] Ara S, Das A, Dey A. Malignant and benign breast cancer classification using machine learning algorithms. In: 2021 International Conference on Artificial Intelligence (ICAI). IEEE; 2021. p.97–101.

[35] Sengar PP, Gaikwad MJ, Nagdive AS. Comparative study of machine learning algorithms for breast cancer prediction. In: 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT). IEEE; 2020. p.796–801.

[36] Ak MF. A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications. Healthcare. 2020;8:111.

[37] Abdar M, Zomorodi-Moghadam M, Zhou X, Gururajan R, Tao X, Barua PD, et al. A new nested ensemble technique for automated diagnosis of breast cancer. Pattern Recognit Lett. 2020;132:123–31.