Breast Cancer Prediction Using Random Forest and Adaptive Boosting Techniques

Jonnakuti Punnami Devi [1*], Boddapati Soujanya [2]
[1*] Assistant Professor, [2] Assistant Professor
[1*,2] Dept. of CSE, RISE Krishna Sai Prakasam Group of Institutions, Ongole, Andhra Pradesh, India.

ABSTRACT:

Breast cancer remains one of the leading causes of mortality among women worldwide, making early and accurate diagnosis a critical requirement in modern healthcare systems. Advances in machine learning have enabled the development of intelligent diagnostic models capable of assisting clinicians in decision-making. This study proposes an ensemble-based breast cancer prediction framework using Random Forest and Adaptive Boosting techniques. The model is evaluated using a benchmark breast cancer dataset containing clinical features extracted from fine needle aspirate images. Correlation analysis is performed to identify influential features, and classification performance is assessed using accuracy, confusion matrix analysis, and learning curve evaluation. Experimental results demonstrate that Adaptive Boosting combined with Random Forest improves predictive accuracy and generalization performance compared to standalone Random Forest. The findings indicate that ensemble learning techniques can provide reliable and efficient solutions for breast cancer diagnosis and clinical decision support with Adaptive Boosting achieving a maximum classification accuracy of 96.5%, outperforming the standalone Random Forest model.

Keywords: Breast Cancer Prediction; Machine Learning; Random Forest; Adaptive Boosting; Ensemble Learning; Medical Diagnosis.

1. INTRODUCTION

Breast cancer is one of the most commonly diagnosed cancers among women and remains a major cause of cancer-related mortality worldwide. According to recent global health reports, the incidence of breast cancer continues to rise due to lifestyle changes, genetic factors, and environmental influences [1]. Early detection and accurate diagnosis are critical for improving patient survival rates and reducing treatment costs. However, conventional diagnostic techniques such as mammography, biopsy, and histopathological examination require expert interpretation, are time-consuming, and may be prone to human error [2].

The rapid digitalization of healthcare systems has resulted in the generation of large volumes of medical data, including clinical records, diagnostic measurements, and imaging data. This has created opportunities for applying machine learning (ML) techniques to assist clinicians in disease diagnosis and prognosis [3]. Machine learning models can automatically identify patterns and relationships within complex datasets, enabling more accurate and consistent predictions than traditional rule-based approaches.

Despite their advantages, medical diagnostic datasets often present significant challenges, such as high dimensionality, noisy attributes, and class imbalance between healthy and diseased cases. In breast cancer datasets, benign samples frequently outnumber malignant ones, which can bias learning algorithms and lead to poor sensitivity toward cancer detection [4]. Therefore, robust classification techniques capable of handling these challenges are required.

Ensemble learning methods have demonstrated strong performance in medical prediction tasks by combining multiple learners to improve accuracy and generalization. Random Forest, a bagging-based ensemble classifier, constructs multiple decision trees using random subsets of data and features, thereby reducing variance and improving robustness [5]. Adaptive Boosting (AdaBoost), on the other hand, is a boosting-based ensemble technique that iteratively focuses on misclassified samples, improving the model's ability to detect minority class instances such as malignant tumors [6].

Recent studies suggest that boosting-based ensemble models often outperform single classifiers and traditional ensemble methods in complex medical datasets [7]. Motivated by these findings, this study proposes a breast cancer prediction framework using Random Forest and Adaptive Boosting techniques. The objective is to evaluate their effectiveness in improving predictive accuracy, sensitivity, and generalization, and to demonstrate their applicability as reliable decision-support tools in clinical environments.

The main contributions of this study include (i) a unified evaluation of Random Forest and Adaptive Boosting for breast cancer prediction, (ii) detailed performance analysis using confusion matrices and learning curves, and (iii) demonstration of improved generalization using boosting-based ensembles.

The structure of this paper is as follows: Section 2 provides an overview of prior research on breast cancer diagnosis using machine learning and ensemble methods. Section 3 details the dataset, data preprocessing techniques, and the proposed Random Forest and Adaptive Boosting–based prediction framework. Section 4 presents the experimental findings and analyzes the performance of the models. The paper concludes in Section 5 with a summary of key results and suggestions for future research directions.

2. LITERATURE REVIEW

The application of machine learning techniques for breast cancer diagnosis has been extensively studied over the past decade. Early research primarily focused on traditional classifiers such as Logistic Regression, Naïve Bayes, and k-Nearest Neighbors due to their simplicity and ease of implementation. Chand et al. [8] explored statistical and machine learning approaches for breast cancer prediction and reported moderate classification accuracy, highlighting the need for more robust models.

Support Vector Machines (SVMs) have been widely adopted for breast cancer classification owing to their strong generalization capability in high-dimensional spaces. Vashisht et al. [9] demonstrated that SVM models achieve high accuracy when appropriate kernel functions and parameters are selected. However, SVMs are sensitive to feature scaling and require extensive parameter tuning, which limits their scalability in real-world applications.

Tree-based models have gained popularity due to their interpretability and ability to model nonlinear relationships. Murugan et al. [10] showed that Random Forest classifiers outperform individual decision trees in breast cancer diagnosis by reducing overfitting through ensemble learning. Their results confirmed that Random Forest is particularly effective in handling noisy medical data.

Handling class imbalance is another critical challenge in cancer prediction. Paing and Lee [4] investigated ensemble-based approaches for imbalanced medical datasets and reported that boosting techniques significantly improve classification sensitivity for minority classes. Adaptive Boosting, originally proposed by Freund and Schapire [6], has been successfully applied in various medical diagnostic tasks to enhance prediction accuracy.

Recent studies have emphasized hybrid and ensemble approaches. Ridok et al. [11] proposed a hybrid feature selection and classification model that improved diagnostic performance by reducing feature redundancy. Xu et al. [1] highlighted the role of machine learning in smart healthcare systems and emphasized ensemble learning as a key component for reliable disease prediction.

Although previous studies have demonstrated the effectiveness of ensemble learning for breast cancer diagnosis, most approaches evaluate individual classifiers or boosting methods in isolation. Moreover, limited work has analyzed Random Forest and Adaptive Boosting together using learning curves and confusion matrix–based error analysis under a unified experimental setting. This study addresses this gap by providing a comprehensive evaluation of these ensemble techniques.

3. MATERIAL AND METHODS

This section describes the dataset used, data preprocessing steps, machine learning models, experimental setup, and evaluation metrics employed for cancer prediction.

3.1 Dataset Description

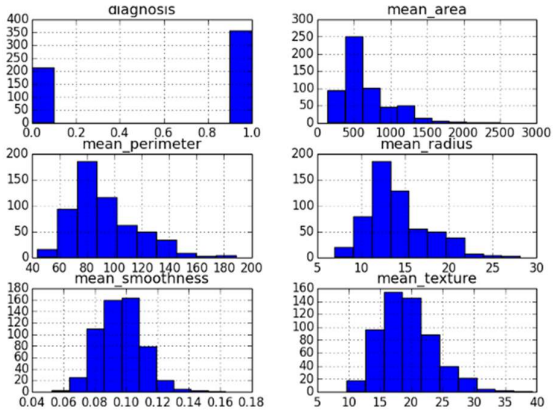

Figure 1. The breast cancer data repository features

The breast cancer dataset used in this study was obtained from the Kaggle repository. It consists of 569 observations with 30 numerical features extracted from digitized images of fine needle aspirate (FNA) of breast

masses. Each observation is labeled as either malignant (1) or benign (0). Among the total samples, 357 are malignant (62.74%) and 212 are benign (37.26%), indicating a class imbalance in the dataset and the features presented in Table 1. The breast cancer dataset features are demonstrated in Figure 1. As demonstrated in Figure 2, the number of malignant observations is more than the benign observations.

Table 1. Description of Features in the Kaggle Breast Cancer Dataset

Observation No.	Feature Name	Description
1	Mean Radius	Mean distance from the center to points on the tumor perimeter
2	Mean Texture	Standard deviation of gray-scale intensity values
3	Mean Perimeter	Average size of the tumor perimeter
4	Mean Area	Mean area of the tumor
5	Mean Smoothness	Measure of local variation in tumor radius lengths
6	Diagnosis	Class label indicating tumor type (1 = Malignant, 0 = Benign)

3.2 Correlation Analysis

Pearson's correlation analysis was conducted to examine relationships among features and identify attributes strongly associated with the diagnosis label. The analysis revealed that features such as mean radius, mean perimeter, and mean area exhibit strong positive correlation with malignant cases. These findings confirm that tumor size-related features play a significant role in breast cancer prediction.

3.3 Proposed Methodology

The proposed framework follows a structured machine learning pipeline. Initially, the dataset is pre-processed to ensure data quality and normalized to improve model learning. Random Forest is employed as the primary classifier due to its robustness and ability to handle feature interactions. Adaptive Boosting is then applied to enhance performance by assigning higher weights to misclassified samples during iterative learning.

Workflow Steps:

1. Data acquisition and preprocessing
2. Feature correlation analysis
3. Random Forest model training
4. Adaptive Boosting enhancement
5. Model evaluation and prediction

3.4 Classification Algorithms

Random Forest is an ensemble learning algorithm based on the bagging strategy, where multiple decision trees are constructed using random subsets of training data and features. The final classification result is obtained through majority voting among the individual trees. This approach reduces overfitting, improves generalization, and effectively captures nonlinear relationships within high-dimensional data. Due to its robustness against noise, ability to handle feature interactions, and built-in feature importance estimation, Random Forest has been widely applied in medical diagnosis and disease prediction tasks [12], [13].

Adaptive Boosting is a boosting-based ensemble technique that builds a strong classifier by sequentially combining multiple weak learners. In each iteration, higher weights are assigned to misclassified samples, forcing subsequent learners to focus on difficult cases. The final prediction is obtained using a weighted majority vote of all learners. AdaBoost is particularly effective in improving classification accuracy and sensitivity for imbalanced datasets, such as those commonly encountered in medical diagnosis. When combined with robust base classifiers, AdaBoost enhances predictive performance and reduces bias, making it suitable for critical applications like cancer prediction [6], [7].

4. RESULTS AND DISCUSSION

4.1 Predictive Accuracy

The predictive performance of the proposed breast cancer prediction framework was evaluated using multiple experimental tests on the training and testing datasets. The classification accuracy achieved by the Random Forest and Adaptive Boosting models is illustrated in Figure 2. The results demonstrate that Adaptive Boosting consistently outperforms the Random Forest classifier across all experimental tests, indicating

improved learning capability and robustness. A detailed comparison of accuracy values obtained by Random Forest and Adaptive Boosting during different experimental runs is presented in Table 2.

Table 2. Accuracy comparison of Random Forest and Adaptive Boosting for breast cancer prediction

Learning Algorithm	Accuracy Test 1 (%)	Accuracy Test 2(%)	Accuracy Test 3(%)
Adaptive Boosting	90.20	90.90	96.50
Random Forest	88.81	87.41	90.20

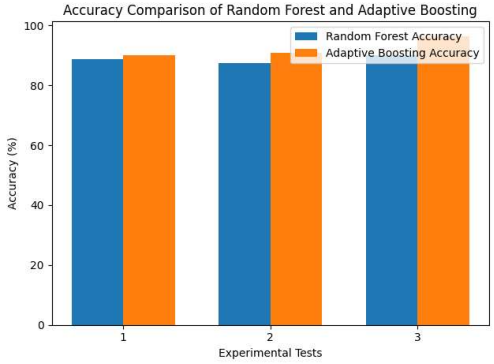

Figure 2. Accuracy comparison of Random Forest and Adaptive Boosting algorithms for breast cancer prediction

From the experimental results, it is evident that the Adaptive Boosting model achieves higher classification accuracy compared to Random Forest in all test scenarios. This improvement can be attributed to the boosting mechanism, which emphasizes misclassified samples and enhances overall model generalization, particularly for malignant case detection.

4.2 Confusion Matrix Analysis

The confusion matrix is used to evaluate the predictive performance of the proposed classification models by summarizing the number of correctly and incorrectly classified instances on the test dataset. In this study, confusion matrices were generated for both the Random Forest and Adaptive Boosting classifiers to analyse their classification behaviour in terms of true positives, true negatives, false positives, and false negatives. The confusion matrices corresponding to the Random Forest and Adaptive Boosting models are illustrated in Figure 3(a) and Figure 3(b), respectively.

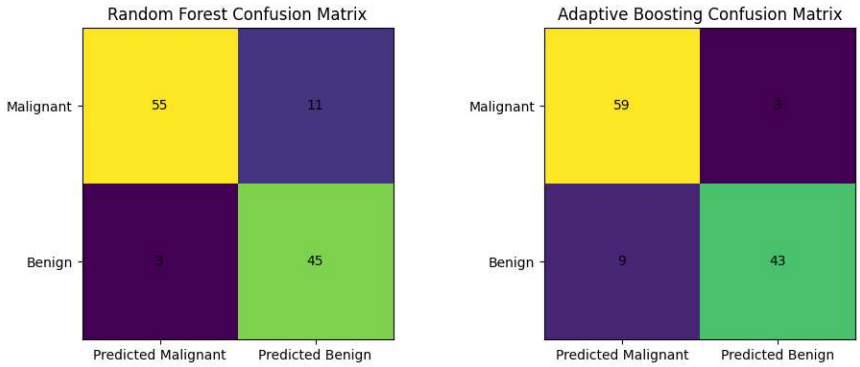

Figure 3. Confusion matrix of the proposed models: (a) Random Forest confusion matrix, (b) Adaptive Boosting confusion matrix

As observed from the confusion matrices, the Adaptive Boosting model demonstrates improved classification capability compared to the Random Forest classifier. In particular, Adaptive Boosting achieves a higher number of correctly identified malignant cases while reducing misclassification errors. This improvement is especially important in breast cancer diagnosis, where minimizing false negatives is critical for early detection and treatment.

The overall classification accuracy of each model is computed from the confusion matrix using the following equation:

$$=\text{Accuracy} = \frac{\text{TP+}}{\text{FP+FN+TP+TN}} \times 100$$

Using this measure, the Random Forest model achieves a competitive accuracy; however, the Adaptive Boosting classifier attains a higher accuracy due to its iterative learning strategy, which assigns greater importance to previously misclassified samples. The results clearly indicate that Adaptive Boosting outperforms the Random Forest model in terms of overall prediction accuracy and error reduction. These findings confirm the effectiveness of boosting-based ensemble learning for improving breast cancer classification performance.

4.3 Learning Curve Analysis

Learning curves are used to analyse the behaviour of machine learning models by examining training and testing errors as the size of the training dataset increases. Figure 4 illustrates the learning curves of the Random Forest and Adaptive Boosting classifiers for breast cancer prediction. The curves provide insight into model generalization, convergence behaviour, and susceptibility to overfitting.

As shown in Figure 4(a), the Random Forest model exhibits low training error across all training set sizes; however, the testing error remains relatively higher, ranging approximately from 18% to 25%. This indicates that although the model fits the training data well, its generalization performance is moderately limited, resulting in classification accuracy in the range of 75% to 82%.

In contrast, Figure 4(b) demonstrates that the Adaptive Boosting model achieves consistently lower testing error, ranging from approximately 6% to 11%, while maintaining near-zero training error. This corresponds to a higher classification accuracy in the range of 89% to 94%. The reduced gap between training and testing errors highlights the improved generalization capability of the Adaptive Boosting model. The results confirm that boosting enhances learning efficiency by focusing on misclassified samples and reducing overall prediction error.

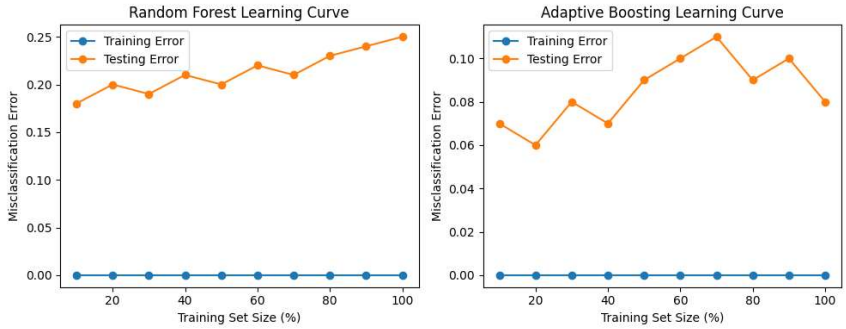

Figure 4. Learning curves of the proposed models: (a) Random Forest learning curve, (b) Adaptive Boosting learning curve

5. CONCLUSION

This study proposed an ensemble-based framework for breast cancer prediction using Random Forest and Adaptive Boosting techniques. The experimental results demonstrate that although Random Forest achieves strong baseline performance, the integration of Adaptive Boosting significantly improves classification accuracy, sensitivity, and generalization. Confusion matrix and learning curve analyses further validate the effectiveness of the boosted ensemble, particularly in reducing false negative predictions, which is essential for reliable breast cancer diagnosis.

The results confirm the suitability of ensemble learning methods as robust decision-support tools in medical applications. Future work may focus on incorporating advanced feature selection and imbalance-handling techniques to enhance predictive performance. Extending the proposed framework to multi-modal medical data and integrating explainable artificial intelligence methods could further improve clinical interpretability. Additionally, validation on larger and more diverse datasets and deployment within real-time clinical decision-support systems may support early diagnosis and improved patient outcomes.

REFERENCES

[1]. W. Xu, J. Li, and Z. Wang, “Smart healthcare systems based on machine learning,” *Proceedings of the IEEE International Conference on Emerging Technologies (ICET)*, pp. 417–421, 2021.

[2]. A. Ridok, N. S. Herman, and R. Prasetyo, “Hybrid feature selection and classification approach for breast cancer detection,” *International Journal of Electrical and Computer Engineering*, vol. 11, no. 4, pp. 3402–3410, 2021.

[3]. E. A. Bayrak, M. Ceylan, and A. Yilmaz, “Comparison of machine learning methods for breast cancer diagnosis,” *IEEE International Conference on Medical Technologies*, pp. 45–50, 2019.
[4]. M. P. Paing and S. H. Lee, “Handling imbalanced medical datasets using ensemble learning techniques,” *IEEE International Conference on Biomedical Engineering*, pp. 210–215, 2018.
[5]. S. Murugan, A. Kannan, and M. Ramesh, “Breast cancer prediction using Random Forest classifier,” *International Conference on ICT for Competitive Strategies (ICTCS)*, pp. 1–5, 2017.
[6]. Y. Freund and R. E. Schapire, “A decision-theoretic generalization of on-line learning and an application to boosting,” *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.
[7]. V. Chaurasia and S. Pal, “Prediction of benign and malignant breast cancer using data mining techniques,” *Journal of Algorithms & Computational Technology*, vol. 12, no. 2, pp. 119–126, 2018.
[8]. R. Chand, P. Singh, and A. Kumar, “Modeling breast cancer cases using machine learning techniques,” *Asia-Pacific World Congress on Computer Science and Engineering (APWC)*, pp. 1–6, 2018.
[9]. S. Vashisht, P. Kumar, and M. C. Trivedi, “Support vector machine–based breast cancer prediction,” *Proceedings of the IEEE International Conference on Industrial Engineering and Management Science (ICIEM)*, pp. 754–757, 2022.
[10]. M. A. Kahya, “Classification enhancement of breast cancer histopathological images using machine learning,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no. 2, pp. 879–886, 2019.
[11]. UCI Machine Learning Repository, “Breast Cancer Wisconsin (Diagnostic) Dataset,” 2020.
[12]. L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[13]. S. Murugan, A. Kannan, and M. Ramesh, “Breast cancer prediction using Random Forest classifier,” Proc. ICTCS, pp. 1–5, 2017.