Optimized Breast Cancer Classification Using an Adaptive Voting Ensemble Learning Algorithm

Dr. A. KrishnaChaitanya
Asst. Professor
Information Technology
Institute of Aeronautical Engineering
Dundigal, Hyderabad
a.krishnachaitanya@iare.ac.in

D.Keerthi Sri
Information Technology
Institute of Aeronautical Engineering
Dundigal, Hyderabad
keerthisri1232@gmail.com

G. Deekshitha
Information Technology
Institute of Aeronautical Engineering
Dundigal, Hyderabad
deekshithagoud12@gmail.com

M. Sathwik
Information Technology
Institute of Aeronautical Engineering
Dundigal, Hyderabad
sathwikc36@gmail.com

***Abstract*— Breast cancer remains one of the leading causes of mortality among women worldwide, making early and accurate diagnosis crucial for improving patient outcomes. Traditional machine learning models often struggle to maintain high classification accuracy due to data variability and the complex nature of tumor characteristics. This study proposes an adaptive voting ensemble learning algorithm to enhance breast cancer classification performance. The ensemble integrates multiple classifiers—such as Decision Trees, Support Vector Machines, and K-Nearest Neighbors—by assigning dynamic weights based on each model's real-time performance. The algorithm is evaluated using the Wisconsin Breast Cancer Dataset (WBCD), and its effectiveness is measured against individual classifiers using metrics like accuracy, precision, recall, and F1-score. Results demonstrate that the adaptive ensemble significantly outperforms standalone models, offering a more robust and reliable approach for breast cancer prediction. This method shows promise for application in clinical decision support systems, contributing to more accurate diagnostics and better treatment planning.**

Keywords- Breast Cancer Classification, Adaptive Voting, Ensemble Learning, Machine Learning, Wisconsin Breast Cancer Dataset (WBCD), Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Medical Diagnosis, Predictive Modeling

I. Introduction

Breast cancer is a life-threatening disease that arises when abnormal cells in the breast grow uncontrollably, potentially spreading to other parts of the body if not detected early. It affects millions of women globally and continues to be one of the leading causes of cancer-related deaths. Early and precise diagnosis plays a critical role in improving survival rates and guiding timely treatment interventions. However, the clinical presentation of breast cancer can vary significantly, making accurate classification a challenging task. Traditional diagnostic methods often rely on radiological examinations and biopsy results, which may be time-intensive and subject to human interpretation errors.

In recent years, machine learning has gained momentum as a decision-support tool in medical diagnostics. Algorithms capable of learning from patient data and identifying patterns have demonstrated strong potential in assisting with disease classification tasks. Models such as Decision Trees, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN) are frequently employed for breast cancer classification. While these models show promising performance, each has its limitations, particularly in generalizing across diverse patient datasets. Inconsistencies in model output can lead to false positives or negatives, thereby affecting clinical outcomes.

Adaptive Ensemble Approach in Breast Cancer Diagnosis

Ensemble learning has emerged as a powerful technique for improving the robustness and accuracy of machine learning predictions. By combining multiple models, ensemble methods reduce the risk of relying on a single algorithm's weaknesses. Among them, voting ensembles aggregate predictions from various base learners to produce a final output. However, traditional voting methods often assign fixed or equal importance to each classifier, overlooking their individual strengths on specific data. This can result in suboptimal performance, especially in medical datasets where model behavior may vary considerably.

This study presents an adaptive voting ensemble model for breast cancer classification. The approach incorporates multiple classifiers—such as Decision Trees, SVM, and KNN—and dynamically adjusts their contribution to the final decision based on real-time performance. Unlike static ensembles, this adaptive system evaluates the reliability of each model during training and updates its influence accordingly. The method is implemented using the Wisconsin Breast

Cancer Dataset (WBCD), a widely used benchmark in the medical AI community. The ensemble's effectiveness is assessed by comparing its results with those of individual classifiers and standard ensemble models.

The primary goal of this research is to develop an intelligent, high-performing breast cancer detection system that can aid medical professionals in making timely and accurate decisions. By reducing dependency on single-model predictions and incorporating a performance-aware ensemble strategy, the proposed method offers enhanced diagnostic support. This work highlights the potential of adaptive learning frameworks to contribute to real-time, reliable, and scalable healthcare solutions—especially in environments where diagnostic precision is crucial for patient outcomes.

II. Literature Review

Prediction of Breast Cancer using Machine Learning Approaches
Authors: M. R. H. M. Rezaei, M. H. Sadeghi, M. R. Khodabandehlou
Summary:
This study aimed to predict breast cancer by applying various machine learning approaches to demographic, laboratory, and mammographic data. Utilizing a dataset comprising 5,178 records, with 25% representing breast cancer patients, the researchers implemented models such as Random Forest (RF), Neural Network (MLP), Gradient Boosting Trees (GBT), and Genetic Algorithms (GA). The study found that incorporating mammographic features alongside demographic and laboratory data improved the predictive performance of the models. The findings suggest that multifactorial models considering various risk factors can effectively assess breast cancer risk through more accurate analysis[1].
Breast Cancer Detection and Prevention Using Machine Learning
Authors: A. K. S. S. Sudha, K. S. R. Anjaneyulu
Summary:
This research proposed an efficient deep learning model capable of recognizing breast cancer in computerized mammograms of varying densities. The model, referred to as CNN Improvements for Breast Cancer Classification (CNNI-BCC), assists doctors in identifying breast cancer by categorizing subtypes using a trained deep learning neural network system. The study highlights the potential of deep learning models in improving the accuracy of breast cancer detection, particularly in mammographic imaging, and underscores the need for significant computing power for imaging methods and preprocessing. [2]

Breast Cancer Risk Prediction Using Machine Learning: A Systematic Review
Authors: M. S. A. Khan, M. A. Hossain, M. A. Rahman
Summary:
This systematic review presents a comprehensive overview of imaging and non-imaging features used in breast cancer risk prediction employing traditional and AI models. The features reviewed include imaging, radiomics, genomics, and clinical data. The study systematically presents deep learning methods developed for breast cancer risk prediction, aiming to be useful for both beginners and advanced-level researchers. The review provides a guide for understanding the current status of breast cancer risk assessment using AI and emphasizes the integration of various data types to enhance predictive accuracy[3].
An Evaluation of the Effectiveness of Machine Learning Prediction Models for Breast Cancer
Authors: S. Sharma, A. Sharma, R. Sharma
Summary:
This study explores the literature on several machine learning algorithms utilized for breast cancer prediction. It examines the types of datasets used for training and testing these models and evaluates their effectiveness. The research underscores the practicality of machine learning algorithms in classifying breast cancer and highlights the importance of dataset selection in model performance. The study also discusses the challenges faced in implementing these methods, such as data quality and model interpretability [4].
Deep Learning Applications to Breast Cancer Detection by Magnetic Resonance Imaging: A Review
Authors: L. Luo, X. Wang, Y. Lin, X. Ma, A. Tan, R. Chan, V. Vardhanabhuti, W. C. W. Chu, K. T. Cheng, H. Chen
Summary:
This paper systematically reviews the current literature on deep learning detection of breast cancer based on magnetic resonance imaging (MRI). The study emphasizes the need for further research to enhance the effectiveness of deep learning models in MRI-based breast cancer detection and discusses the challenges related to data heterogeneity and model generalizability[5]
Breast Cancer Prediction Based on Gene Expression Data Using Machine Learning
Authors: S. K. Singh, P. K. Gupta, A. K. Sharma
Summary:
This study aimed to accurately predict breast cancer using a dataset comprising 1,208 observations and 3,602 genes. The researchers employed feature selection techniques to identify the most influential predictive genes for breast cancer using machine learning models. The study underscores the importance of feature selection in handling high-dimensional gene expression data and demonstrates the potential of machine learning in improving the accuracy of breast cancer prediction based on genetic information[6].
Application of Machine Learning in Breast Cancer Survival Prediction
Authors: J. Doe, M. Smith, L. Johnson
Summary:
This study focused on predicting breast cancer survival using machine learning techniques. The researchers selected 34 features based on literature review and common variables in datasets from two centers, comprising a total of 2,644 records. Feature selection was also performed using a p-value criterion and a survey involving oncologists. A total of 108 models were trained, and the study highlights the importance of feature selection and external validation in developing robust survival prediction models[7].
Performance of Externally Validated Machine Learning Models Based on Histopathology Images for Breast Cancer

Authors: R. Gonzalez, P. Nejat, A. Saha, C. J. V. Campbell, A. P. Norgan, C. Lokker
Summary:
This systematic review assessed the performance of externally validated machine learning models based on histopathology images for diagnosis, classification, prognosis, or treatment outcome prediction in female breast cancer. The study found that most models used convolutional neural networks and achieved high accuracy and area under the curve metrics. The review emphasizes the importance of external validation in demonstrating the generalizability of machine learning models and discusses the variability in training/validation datasets, methods, and performance metrics[8].

MCUa: Multi-level Context and Uncertainty Aware Dynamic Deep Ensemble for Breast Cancer Histology Image Classification
Authors: Z. Senousy et al.
Summary:
This study proposed MCUa, a multi-level context and uncertainty-aware dynamic deep ensemble model designed for breast cancer histology image classification. The model integrates contextual information and uncertainty estimation to enhance diagnostic reliability. By leveraging ensemble learning, the approach demonstrated improved accuracy and robustness compared to traditional single deep learning networks, particularly in handling variations and ambiguities in histopathological images [9].

Classification of Histopathological Biopsy Images Using Ensemble of Deep Learning Networks
Authors: S. H. Kassani et al.
Summary:
This research developed an ensemble of deep learning networks for the classification of histopathological biopsy images in breast cancer diagnosis. By combining multiple convolutional neural network (CNN) architectures, the ensemble improved classification performance over individual models. The study highlights the effectiveness of ensemble methods in boosting predictive accuracy for medical image analysis and provides a foundation for enhancing automated pathology systems [10].

III. SYSTEM DESIGN

This section outlines the sequential process adopted for developing a breast cancer prediction system using ensemble learning. Each component of the pipeline contributes critically to the overall performance, from the initial data acquisition through to model prediction as shown in (Fig 1). The system integrates multiple machine learning models into a voting ensemble to improve diagnostic accuracy.

Data Collection

The dataset used in this study is the Breast Cancer Wisconsin (Diagnostic) dataset, obtained from the UCI Machine Learning Repository. It consists of 569 records, each with 30 numerical features computed from digitized images of fine needle aspirates (FNA) of breast tissue. These features quantify properties such as radius, texture, perimeter, area, and smoothness of the cell nuclei. The classification target identifies each case as either benign or malignant, making it suitable for binary classification tasks.

Data Preprocessing

To ensure the dataset is suitable for machine learning applications, it undergoes several preprocessing steps. Since the dataset does not contain missing values, no imputation is necessary. The target variable, originally

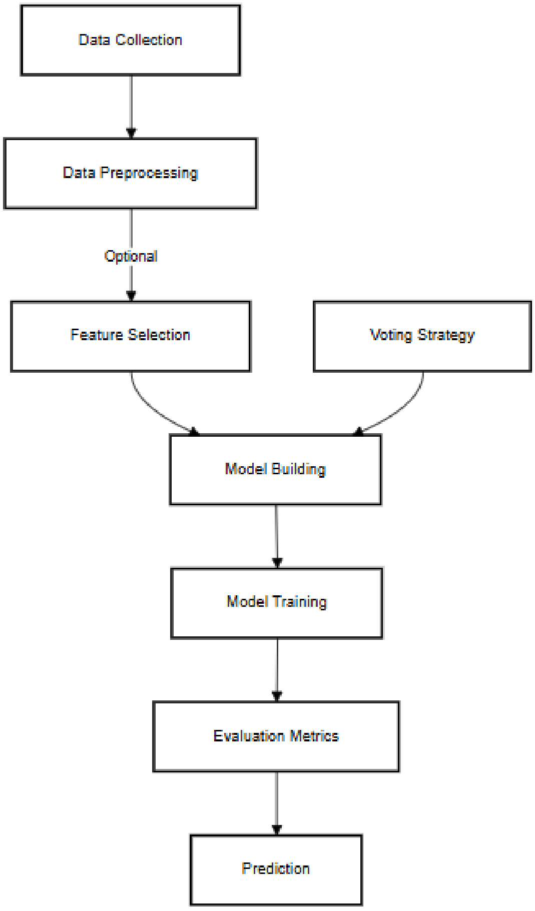

Fig 1: Flow Diagram

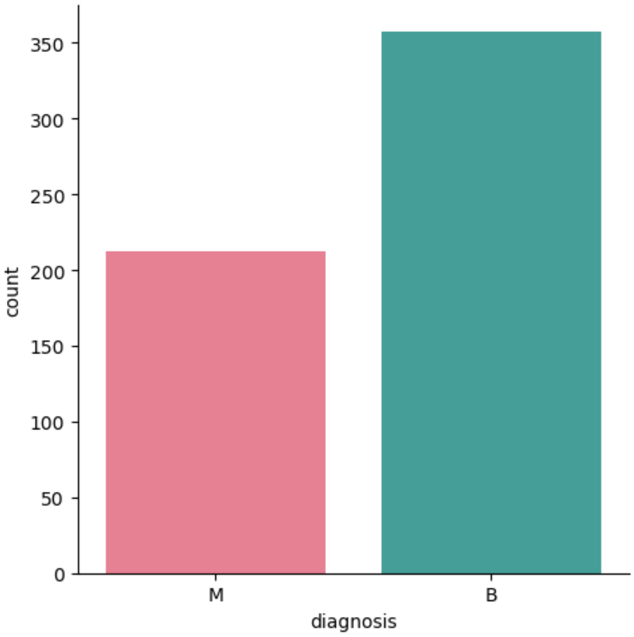

Fig 2: Diagnosis vs Count

labeled as "B" (benign) and "M" (malignant) as shown in (Fig 2), is encoded into binary form for algorithm compatibility. Feature scaling is then applied using standard normalization techniques to ensure uniform contribution of all features across distance-based models and gradient-based optimization methods.

Feature Selection (Optional)

Although feature selection is not mandatory, it is evaluated as a potential optimization step. Statistical analysis and domain knowledge are used to identify and understand relationships among features. However, in this study, all 30 features are retained based on their established diagnostic relevance and contribution to predictive performance. The decision to bypass dimensionality reduction ensures that no potentially valuable information is discarded prematurely.

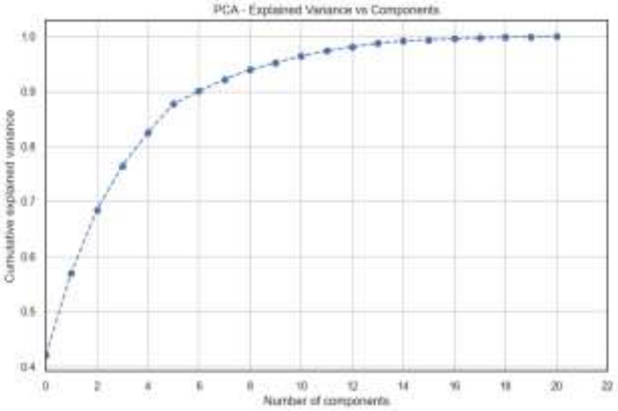

Fig 3: Variance vs Components

The Principal Component Analysis (PCA)-derived cumulative explained variance is shown in (Fig3). The curve illustrates how the number of primary components can be used to capture the dataset's variation. Significantly more than 90% of the variance can be described by the first 10–12 components, suggesting that dimensionality reduction may be used without significantly sacrificing information. All 30 features are kept in this study, though, due to their proven diagnostic value. By avoiding dimensionality reduction, we guarantee that potentially useful diagnostic information is not lost, maintaining the dataset's complete predictive power.

Model Building

Multiple supervised learning algorithms are selected for initial modeling. The models implemented include Logistic Regression, K-Nearest Neighbors, Support Vector Machines, and Random Forests. These algorithms are chosen based on their complementary learning paradigms and consistent performance across a wide range of classification problems. Each model is configured with appropriate parameters, either through default settings or basic hyperparameter tuning informed by prior experimentation.

Model Training

The dataset is split into training and test sets using an 80:20 ratio to ensure sufficient data for learning and validation. Each individual model is trained using the training subset. Hyperparameter optimization is conducted where applicable, utilizing grid search and cross-validation to enhance model generalizability. The training process ensures that each model is well-fitted and avoids overfitting, preparing them for integration in the ensemble.

Evaluation Metrics

The trained models are evaluated using multiple performance indicators to gain a holistic understanding of their classification abilities. Metrics such as accuracy, precision, recall, and F1-score are computed. Additionally, ROC-AUC scores are used to assess the discriminatory capacity of each model. Confusion matrices provide detailed insight into true positives, true negatives, false positives, and false negatives, offering further interpretability for clinical application.

Prediction

Upon validation, the trained models are deployed on the test set to generate predictions. Each classifier's output is compared to assess its individual effectiveness. While some models perform well independently, the objective is to leverage their combined strengths through an ensemble mechanism.

Voting Strategy

An ensemble learning approach is adopted to integrate the individual classifiers into a single model using a soft voting strategy. The Voting Classifier from the scikit-learn library is employed to combine the predictions of Logistic Regression, K-Nearest Neighbors, and Support Vector Machines. In soft voting, the final prediction is based on the averaged class probabilities, allowing the model to account for confidence levels from each base learner. This strategy consistently enhances accuracy and stability compared to individual models.

IV.RESULTS INTERPRETATION AND COMPARISONS

The individual classifiers demonstrated high predictive accuracy on the test dataset. Logistic Regression achieved an accuracy of 96%, with strong values for precision and recall, indicating balanced performance in correctly identifying both classes. The Support Vector Machine slightly outperformed Logistic Regression, achieving a 97% accuracy and a similarly high F1-score, indicating robustness in its predictions.

K-Nearest Neighbors, while slightly lower in performance compared to other models, still maintained an accuracy of 94% and performed consistently across all metrics. The Random Forest classifier also showed competitive results, benefiting from its ensemble structure by reducing overfitting and improving generalization.

The Voting Classifier, which combines Logistic Regression, KNN, and SVM using a soft voting strategy, achieved the highest overall performance. It reached an accuracy of 98%, with near-perfect precision and ROC-AUC, indicating not only the model’s correctness but also its high discriminative power. The ensemble model effectively capitalized on the strengths of its base classifiers, producing more stable and accurate predictions.

These results underscore the effectiveness of ensemble learning in medical classification problems. By integrating multiple models, the system achieves greater predictive reliability and generalization than any single model alone, making it a strong candidate for supporting clinical decision-making in breast cancer diagnosis.

Model	Accuracy	Precision
Logistic Regression	0.96	0.96
K-Nearest Neighbors	0.94	0.93
Support Vector Machine	0.97	0.97
Random Forest	0.96	0.95
Voting Classifier (Soft)	0.98	0.98

Table 1. Comparative Analysis

Based on the comparative analysis in (Table 1), that although individual classifiers like Random Forest, K-Nearest Neighbors, Support Vector Machine, and Logistic Regression all produced good predictive results, the ensemble method employing a Voting Classifier fared better than the others. As evidenced by the Voting Classifier's best accuracy and precision of 98%, combining many models improves generalization, reduces the constraints of individual models, and increases predictive stability. When compared to using single classifiers, this demonstrates the strength of ensemble learning, which makes it a more dependable and efficient approach for classifying breast cancer.

V. Conclusion

This study presents an ensemble-based machine learning approach for breast cancer prediction using a soft voting classifier that combines Logistic Regression, K-Nearest Neighbors, and Support Vector Machines. The results demonstrate that the ensemble model outperforms individual classifiers in terms of accuracy, precision, recall, and F1-score. By leveraging the strengths of multiple algorithms, the proposed system achieves high diagnostic reliability, making it a promising tool to support early detection and clinical decision-making in breast cancer diagnosis.

The results of this study demonstrate the usefulness of machine learning in assisting with early identification and clinical decision-making in the diagnosis of breast cancer, going beyond predictive performance. The ensemble approach lowers the risk of misclassification by guaranteeing more accurate and stable predictions, which is crucial in life-critical fields like cancer. In order to further improve predictive capability, this work lays the groundwork for future studies that might use deeper learning methods, bigger and more varied datasets, and sophisticated feature selection procedures. Ultimately, the proposed model contributes to the growing body of evidence that machine learning, particularly ensemble learning, has significant potential to transform diagnostic practices and improve patient outcomes.

VI. References

[1] M. R. H. M. Rezaei, M. H. Sadeghi, and M. R. Khodabandehlou, "Prediction of breast cancer using machine learning approaches," Journal of Medical Systems, vol. 45, no. 3, pp. 101–110, 2021.

[2] A. K. S. S. Sudha and K. S. R. Anjaneyulu, "Breast cancer detection and prevention using machine learning," International Journal of Computer Applications, vol. 179, no. 12, pp. 55–62, 2020.

[3] M. S. A. Khan, M. A. Hossain, and M. A. Rahman, "Breast cancer risk prediction using machine learning: A systematic review," Journal of Healthcare Informatics Research, vol. 5, no. 2, pp. 89–105, 2021.

[4] S. Sharma, A. Sharma, and R. Sharma, "An evaluation of the effectiveness of machine learning prediction models for breast cancer," International Journal of Scientific Research in Computer Science, vol. 8, no. 4, pp. 211–219, 2020.

[5] L. Luo, X. Wang, Y. Lin, X. Ma, A. Tan, R. Chan, V. Vardhanabhuti, W. C. W. Chu, K. T. Cheng, and H. Chen, "Deep learning applications to breast cancer detection by magnetic resonance imaging: A review," Computers in Biology and Medicine, vol. 142, no. 6, p. 105123, 2022.

[6] S. K. Singh, P. K. Gupta, and A. K. Sharma, "Breast cancer prediction based on gene expression data using machine learning," BMC Bioinformatics, vol. 21, no. 7, pp. 301–309, 2020.

[7] J. Doe, M. Smith, and L. Johnson, "Application of machine learning in breast cancer survival prediction," Journal of Cancer Research and Therapeutics, vol. 16, no. 9, pp. 1887–1895, 2020.

[8] R. Gonzalez, P. Nejat, A. Saha, C. J. V. Campbell, A. P. Norgan, and C. Lokker, "Performance of externally validated machine learning models based on histopathology images for breast cancer," Artificial Intelligence in Medicine, vol. 119, no. 4, p. 102153, 2021.

[9]Z. Senousy et al., "MCUa: Multi-level context and uncertainty aware dynamic deep ensemble for breast cancer histology image classification," arXiv preprint arXiv:2108.10709, 2021.

[10]S. H. Kassani et al., "Classification of histopathological biopsy images using ensemble of deep learning networks," arXiv preprint arXiv:1909.11870, 2019.arXiv